SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Sanja Kovačević

METODA UPARIVANJA PO VJEROJATNOSTI SKLONOSTI

Diplomski rad

Voditelj rada:
prof. dr. sc. Sonja Radas

Zagreb, srpanj 2014.

Ovaj diplomski rad obranjen je dana ____________________ pred ispitnim povjerenstvom u sastavu:

1. ____________________________, predsjednik
2. ____________________________, član
3. ____________________________, član

Povjerenstvo je rad ocijenilo ocjenom _______________.

Potpisi članova povjerenstva:

1. ____________________________
2. ____________________________
3. ____________________________

Zahvaljujem se mentorici prof. dr. sc. Sonji Radas na velikoj pomoći i brojnim savjetima koje mi je pružila tijekom izrade diplomskog rada. Najveće hvala mojim roditeljima, Borni i svim mojim prijateljima na razumijevanju i podršci koju su mi pružili tokom studiranja. Ovim putem također želim zahvaliti i svim prijateljima i kolegama koje sam upoznala na fakultetu, što su mi uljepšali moje studentsko razdoblje.

Sadržaj

Sadržaj **iv**

Uvod **4**

1 Procjena prosječnog efekta tretmana **5**
1.1 Uvod u procjenu prosječnog efekta tretmana 5
1.2 Kontračinjenično okruženje i problem samoselekcije 6
1.3 Metode koje pretpostavljaju ”nezavisnost o tretmanu” 8
1.3.1 Regresijske metode 10
1.3.2 Metode bazirane na vjerojatnosti sklonosti 16
1.4 Metode instrumentalnih varijabli 21
1.4.1 Motivacija za korištenje instrumentalnih varijabli 21
1.4.2 2SLS metoda kod višestrukih instrumenata 23
1.4.3 Korištenje instrumentalnih varijabli za procjenu prosječnog efekta tretmana 24
1.4.4 Procjena lokalnog prosječnog efekta tretmana (LATE-a) 33

2 Utjecaj IWT subvencija na R&D u Flandriji **36**
2.1 Value for money? New microeconometric evidence on public R&D grants in Flanders 36
2.2 IWT subvencije 36
2.3 Ekonometrijske metode 37
2.4 Metoda uparivanja 38
2.5 Podaci 39
2.6 Varijable 40
2.6.1 Vremensko usklađivanje varijabli 41
2.7 Deskriptivna statistika 41
2.8 Ekonometrijski rezultati 41
2.8.1 Stabilnost efekta tretmana tokom vremena 43
2.8.2 Učinak više subvencioniranih projekata istovremeno 44

2.8.3 Učinak uzastopnog ponavljanja subvencija na efekt tretmana . . . 46
2.8.4 Učinak primanja subvencija iz drugih izvora na efekt tretmana . . 46
2.8.5 Promatranje inovativnih kompanija zasebno 47
2.8.6 Promatranje KMO programa zasebno 48
2.8.7 Test robusnosti pomoću instrumentalnih varijabli 49

Bibliografija **53**

Uvod

Cilj većine empirijskih istraživanja u ekonomiji i drugim znanostima je kako promjena jedne varijable, na primjer varijable w, utječe na promjenu druge varijable, na primjer y. Mogli bismo promatrati brojne situacije, kao što su: utječe li godina dodatnog školovanja na rast mjesečne plaće, da li povećanje promocije nekog proizvoda utječe na porast prodaje tog proizvoda, ili da li uvođenje obaveznog cijepljenja značajno smanjuje broj oboljelih od neke bolesti.

Kod većine ekonometrijskih studija ono što se procjenjuje ili testira hipotezama je očekivanje varijable y, koju još nazivamo i **regresand-varijabla**, **zavisna varijabla** ili **varijabla odaziva**, uvjetno na skup **nezavisnih varijabli**, **regresorskih varijabli**, **eksplanatornih varijabli**, **kontrolnih varijabli** ili **kovarijabli** koje najčešće označavamo sa $\mathbf{x} = (x_1, x_2, ..., x_K)$. $\mathrm{E}(y \mid \mathbf{x})$ možemo tada zapisati kao $\mathrm{E}(y \mid \mathbf{x}) = \mu(x_1, x_2, ..., x_K) = \mu(\mathbf{x})$. Funkcija $\mu(\mathbf{x})$ određuje kako se prosječna vrijednost y mijenja kada se elementi $\mathbf{x}$ mijenjaju. Ako je broj parametara konačan, govorimo o **parametarskom modelu** $\mathrm{E}(y \mid \mathbf{x})$.

Često korištena pretpostavka u ekonomiji je **"ceteris paribus"** pretpostavka, što znači da se neki od faktora, koji utječu na ono što procjenjujemo, drže fiksnima. Pri procjeni $\mathrm{E}(y \mid w, \mathbf{c})$, vektor $\mathbf{c}$ određuje skup varijabli koje držimo fiksnima dok promatramo utjecaj od w na očekivanu vrijednost y. Razlog za fiksiranje tih varijabli je taj što smatramo da je w koreliran s elementima u $\mathbf{c}$. Ako se podaci za y, w i $\mathbf{c}$ mogu prikupiti iz slučajnog uzorka populacije koju promatramo, $\mathrm{E}(y \mid w, \mathbf{c})$ je lako procijeniti.

Jedan od korisnih "alata" za manipuliranje uvjetnim očekivanjima je **"zakon iteriranih očekivanja"** (*eng. law of iterated expectations = LIE*). Neka je $\mathbf{x}$ slučajni vektor koji je funkcija od $\mathbf{w}$, tj. $\mathbf{x} = \mathrm{f}(\mathbf{w})$. To nam govori da ako je poznat $\mathbf{w}$, poznat nam je i $\mathbf{x}$. Najopćenitije varijante LIE su tada:

$$E(y \mid \mathbf{x}) = E(E(y \mid \mathbf{w}) \mid \mathbf{x}) \,,$$

$$E(y \mid \mathbf{x}) = E(E(y \mid \mathbf{x}) \mid \mathbf{w})$$

i

$$E(y \mid \mathbf{x}) = E(E(y \mid \mathbf{x}, \mathbf{z}) \mid \mathbf{x}) \ .$$

Također vrijedi i:

$$E(y) = E(E(\ y \mid \mathbf{x}))$$

Naime, ako pretpostavimo da $\mathbf{x}$ poprima vrijednosti $\mathbf{c}_1$, $\mathbf{c}_2$, ..., $\mathbf{c}_m$, s vjerojatnostima redom p_1, ..., p_m, onda je:

$$E(y) = p_1 \cdot E(y \mid \mathbf{x} = \mathbf{c}_1) + ... + p_m \cdot E(y \mid \mathbf{x} = \mathbf{c}_m) = E(E(y \mid \mathbf{x}))$$

U ovom radu najviše ćemo se koncentrirati na procjenu prosječnog efekta tretmana, koristeći metode koje pretpostavljaju nezavisnost o tretmanu (regresijske metode i metode bazirane na vjerojatnosti sklonosti) te metode instrumentalnih varijabli. Procjenitelji koje ćemo konstruirati imat će sljedeća svojstva:

Definicija 0.1: Neka je { $\hat{\theta}_n$: n= 1, 2, ... } niz procjenitelja vektora $\theta \in \Theta$. Ako je $\hat{\theta}_n \xrightarrow{p} \theta$, kažemo da je $\hat{\theta}_n$ **konzistentan** procjenitelj za θ.

Definicija 0.2: Procjenitelj $\hat{\theta}$ je **nepristran** procjenitelj za θ ako je E($\hat{\theta}$) = θ.

Definicija 0.3: Neka je { $\hat{\theta}_n$: n= 1, 2, ... } niz procjenitelja vektora $\theta \in \Theta$. Ako je $\sqrt{N}(\hat{\theta}_n - \theta) \xrightarrow{d}$ N($\mathbf{0}$, $\mathbf{V}$), kažemo da je $\hat{\theta}_n$ $\sqrt{N}$**- asimptotski normalno distribuiran**.

Definicija 0.4: Niz { Z_n: n= 1, 2, ... } je **asimptotski normalno distribuiran** ako konvergira po distribuciji slučajnoj varijabli Z ~ N(0,1).

Definicija 0.5: Neka su $\hat{\theta}_1$ i $\hat{\theta}_2$ dva nepristrana procjenitelja za isti nepoznati parametar θ. Kažemo da je procjenitelj $\hat{\theta}_1$ efikasniji od procjenitelja $\hat{\theta}_2$ ako je Var($\hat{\theta}_1$) < Var($\hat{\theta}_2$).

Definicija 0.6: Ako je $\sqrt{N}(\hat{\theta}_n - \theta) \sim$ N($\mathbf{0}$, $\mathbf{V}$) gdje je $\mathbf{V}$ pozitivno definitna s j dijagonalom oblika v_{jj} , a $\mathbf{V}_n$ je konzistentan procjenitelj za $\mathbf{V}$, tada je **asimptotska standardna greška**

od $\hat{\theta}_{nj}$ jednaka $(\hat{v}_{njj}\ /\ n)^{1/2}$.

Asimptotsku standardnu grešku možemo shvatiti kao procjenu standardne devijacije elemenata u $\hat{\theta}_n$.

Definicija 0.7: **”Kitchen sink” regresija** je neformalni naziv za regresiju kod koje se koristi puno eksplanatornih varijabli, i onih bez teorijske ”podloge”, kako bi se eventualno otkrio neki statistički uzorak, tj. odnos između zavisne i nezavisnih varijabli koji se može primijeniti i na drugim skupovima podataka.

Za korištenje metoda instrumentalnih varijabli korisno je prvo definirati pojam i navesti neka osnovna svojstva linearne projekcije:

Definicija 0.8: **Linearna projekcija** y na 1, x_1, x_2, ... , x_k definira se kao:

$$\mathrm{L}(\,y \mid 1, x_1, x_2, \dots, x_k\,) = \mathrm{L}(\,y \mid 1, \mathbf{x}\,) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k = \beta_0 + \mathbf{x}\boldsymbol{\beta}$$

gdje je po definiciji,

$$\boldsymbol{\beta} \equiv [Var(\mathbf{x})]^{-1}\ \mathrm{Cov}(\mathbf{x}, y), \quad \beta_0 \equiv \mathrm{E}(y) - \mathrm{E}(\mathbf{x})\boldsymbol{\beta} = \mathrm{E}(y) - \beta_1\mathrm{E}(x_1) - .... - \beta_k\mathrm{E}(x_k).$$

Sada y možemo zapisati kao: $y = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u$,

gdje je $\mathrm{E}(u^2) < \infty$, E(u)=0 i $\mathrm{Cov}(x_j, u) = 0$, $j = 1, 2, \dots$, k.

Vidimo da je u nekoreliran sa svim elementima $\mathbf{x}$ pa za varijable x_j kažemo da su **egzogene**. Za varijablu koja je korelirana s greškom modela kažemo da je **endogena**.

Kako bismo konzistentno procijenili koeficijente β_j običnom metodom najmanjih kvadrata (**OLS**), nužan uvjet je da je E(u)=0 i $\mathrm{Cov}(x_j, u) = 0$, $j = 1, 2, \dots$, k, što u ovom slučaju vrijedi.

Definicija 0.9: **Binarni model** (*eng. binary response model*) y na $\mathbf{x}$ predstavlja model u kojem se procjenjuje uvjetna vjerojatnost $\mathrm{P}(y = 1 \mid \mathbf{x})$ za binarnu varijablu y i može označavati probit, logit ili linearni vjerojatnosni model.

Vektor **x** označava vektor kovarijabli koje utječu na y. **Probit model** koristi funkciju distribucije standardne normalne slučajne varijable za procjenu. Preciznije,

$$P(y = 1 \mid \mathbf{x}) = F(\mathbf{x}'\beta) = \Phi(\mathbf{x}'\beta)$$

Za procjenu vektora β koristi se **metoda maksimalne vjerodostojnosti**. Maksimizira se funkcija:

$$ln\ L(\beta) = \sum_{i=1}^{n} \Big(y_i \cdot ln\ \Phi(\mathbf{x}_i'\beta) \ + \ (1 - y_i) \cdot ln(1 - \Phi(\mathbf{x}_i'\beta)) \Big)$$

gledano po svim jedinkama i =1,2, .., n.

Logit model na analogan način procjenjuje traženu vjerojatnost. Jedina razlika je što se umjesto funkcije Φ koristi funkcija distribucije F($\mathbf{x}'\beta$) = $e^{\mathbf{x}_i'\beta}$ / (1 + $e^{\mathbf{x}_i'\beta}$).

Poglavlje 1

Procjena prosječnog efekta tretmana

1.1 Uvod u procjenu prosječnog efekta tretmana

Procjena prosječnog efekta tretmana *(eng. average treatment effect - ATE*) važan je dio evaluacije programa, poput evaluacije obuke za posao, vladinih programa poticaja ili nekog medicinskog tretmana. ATE mjeri razliku očekivanih ishoda tretiranih i netretiranih jedinki. Što je ta razlika veća, znači da je efekt tretmana jači. Za svaku jedinku možemo promatrati njezin ishod sa i bez tretmana, no to nije moguće ostvariti istovremeno budući da se jedinke ne mogu istovremeno nalaziti u dva (suprotna) stanja. Ipak, kako bismo bili što bliže toj mogućnosti, postoje metode uparivanja (*eng. matching*) koje pokušavaju pronaći što sličnije jedinke, od kojih je jedna podvrgnuta tretmanu, a druga nije. Jedna od takvih metoda je metoda uparivanja po vjerojatnosti sklonosti (*eng. propensity score matching*) koju su prvi put iznijeli Paul Rosenbaum i Donald Rubin 1983. godine.

Ako se na slučajan način odredi koje jedinke će biti tretirane, a koje ne, tj. ako se na slučajan način jedinke razdijele u te dvije skupine, možemo zaključiti da su za veće grupe te dvije skupine (u prosjeku) identične. Nažalost, u mnogo slučajeva randomizacija tretmana nije moguća budući da (ovisno o potencijalnoj koristi tretmana) i same jedinke sudjeluju u odluci žele li se podvrgnuti tretmanu. Tako, na primjer, ne možemo promatrati utjecaj pohađanja privatnog fakulteta na kasniji dohodak. Naime, odabir fakulteta vrlo vjerojatno nije "slučajan", već ovisi o mnogim faktorima poput cijene školarine, kvalitete nastave i sl.

Osim metoda baziranih na vjerojatnosti sklonosti, za procjenu ATE koriste se i druge metode poput regresijskih metoda i onih baziranih na instrumentalnim varijablama. U ovome poglavlju posebno ćemo objasniti svaku od njih, kao i dati primjere za njihovo korištenje.

1.2 Kontračinjenično okruženje i problem samoselekcije

U modernoj literaturi koja govori o efektima tretmana svaka jedinka (ili agent) ima ishod bez tretmana i onaj s tretmanom. Ako sa y_1 označimo ishod s tretmanom, a sa y_0 ishod bez tretmana, nemoguće je za pojedinca opaziti i y_1 i y_0 budući da se pojedinac ne može istovremeno nalaziti u oba stanja. Pojam "tretman" ima široko značenje i može predstavljati mnogo toga - medicinski tretman, školovanje, vladin program i sl.

Varijable y_1 i y_0 mogu biti binarne (na primjer, kada bismo promatrali primaju li pojedinci socijalnu pomoć ili je li osoba zaposlena), no moguće je i da su neprekidno distribuirane (ako bismo, na primjer, promatrali plaću ili razinu hormona u krvi - ukoliko se radi o medicinskom tretmanu).

Odsada pa nadalje pretpostavljamo da imamo **uzorak** iz populacije koji je **nezavisan** i **jednako distribuiran**. Time smo isključili mogućnost da tretman na jednoj jedinki ima utjecaja na ishod neke druge jedinke. Pretpostavku da tretman na jedinki i ima utjecaja samo na ishod i-te jedinke nazivamo **stable unit treatment value assumption**, skraćeno **(SUTVA)**. Zbog toga što je naš uzorak slučajan, SUTVA vrijedi.

Neka je w binarna varijabla i neka w=1 označava podvrgnutost tretmanu, a w=0 situaciju u kojoj nije bilo primjene tretmana. Uređena trojka (y_0, y_1, w) predstavlja slučajan vektor iz odabrane populacije, dok vektor (y_{i0}, y_{i1}, w_i) pridružujemo i-toj jedinki te populacije.

Kako bismo izmjerili efekt tretmana, zanima nas razlika između ishoda sa i bez tretmana, y_1- y_0. Budući da se radi o slučajnoj varijabli, moramo odrediti koju karakteristiku njezine distribucije želimo procijeniti. Postoji više varijanti odabira:

1. **prosječan efekt tretmana (average treatment effect - ATE)** :

$$ATE \equiv E(y_1 - y_0) \tag{1.1}$$

   ATE predstavlja očekivani efekt tretmana slučajno izabrane osobe iz populacije. Budući da se prosjek računa za cijelu populaciju, uključene su i jedinke koje nikad ne bi bile izabrane za uzorak jer nisu "pogodne" za tretman. Na primjer, testirajući utjecaj obuke za posao na dohodak, bilo bi logično isključiti milijunaše iz ATE jednadžbe, tj. uključiti samo one osobe čiji je dohodak ispod određenog praga.

2. **prosječan efekt tretmana tretiranih jedinki (average treatment effect on the treated -** ATE_1) :

$$ATE_1 \equiv E(y_1 - y_0 \mid w = 1) \tag{1.2}$$

ATE_1 predstavlja prosječan efekt tretmana, ali samo onih jedinki koje su sudjelovale u programu.

3. **lokalni prosječan efekt tretmana (local average treatment effect - LATE)** :

   LATE procjenjuje efekt tretmana koristeći instrumentalne varijable. Njegova prednost je u tome što ga je, i uz slabe pretpostavke, moguće procijeniti. Mane su što se mjeri lokalno, na subpopulaciji koju je općenito teško identificirati, te što ovisi o određenoj instrumentalnoj varijabli koja nam je dostupna. O instrumentalnim varijablama i LATE više ćemo reći kasnije u radu.

Umjesto jednadžbi (1.1) i (1.2) možemo promatrati i njihova uvjetna očekivanja. Neka je x opažena kovarijabla. ATE uvjetno na x je E(y_1 - y_0 | x), a ATE_1 uvjetno na x je E(y_1 - y_0 | x, w=1). Ovisno o odabranom x, možemo definirati ATE za razne podskupove populacije. Tako, na primjer, x može označavati dohodak prije tretmana ili, kao binarna varijabla, označavati rasu, spol ili stupanj siromaštva.

Kao što smo već prije napomenuli, nemoguće je istovremeno opaziti i y_0 i y_1 za pojedinca, pa je time i teško procijeniti jednadžbe (1.1) i (1.2). Ono što opažamo (skupa s varijablom w) je:

$$y = (1 - w)y_0 + wy_1 \tag{1.3}$$

Pretpostavimo da je indikator tretmana w statistički nezavisan od varijabli y_0 i y_1, kao što je slučaj kada tretman slučajno rasporedimo po agentima. Ta činjenica implicira jednakost jednadžbi ATE i ATE_1: E(y_1- y_0 | w=1) = E(y_1- y_0). Štoviše, jednostavno je procijeniti ATE budući da iz jednadžbe (1.3) (uz uvjet da je w=1, pa je i y=y_1) i činjenice da je w nezavisan od y_1 slijedi:

E(y | w=1)= E(y_1 | w=1)= E(y_1)

Slično, vrijedi i:

E(y | w=0)= E(y_0 | w=0)= E(y_0)

Dakle,

$$ATE = ATE_1 = E(\, y_1 - y_0) = E(\, y \mid w = 1) - E(\, y \mid w = 0) \tag{1.4}$$

Desnu stranu jednostavno procijenimo koristeći srednju vrijednost uzorka, točnije razliku srednje vrijednosti y za tretirane jedinke i srednje vrijednosti y za netretirane jedinke. Stoga, randomizacija tretmana po uzorku osigurava da je procjenitelj razlike srednjih vrijednosti *(eng. difference-in-means estimator)* nepristran, konzistentan i asimptotski normalan. Čak i uz slabiji uvjet: E(y_0 | w) = E(y_0) i E(y_1 | w) = E(y_1) će to vrijediti.

Randomizacija tretmana često je neizvediva kod evaluacije programa. U većini slučajeva pojedinci, barem djelomično, određuju hoće li prihvatiti tretman i njihove odluke vezane su uz potencijalne koristi tretmana. Drugim riječima, postoji prisutnost **samoselekcije** *(eng. self-selection).*

Razliku E(y | w=1) - E(y | w=0) možemo prikazati kao:

$$
\begin{aligned}
E(\, y \mid w = 1) - E(\, y \mid w = 0) &= E(\, y_0 \mid w = 1) - E(\, y_0 \mid w = 0) + E(\, y_1 - y_0 \mid w = 1) = \\
&= [\; E(\, y_0 \mid w = 1) - E(\, y_0 \mid w = 0)\;] + ATE_1
\end{aligned}
\tag{1.5}
$$

Ako pretpostavimo da su w i y_0 nezavisni, tj. vrijedi da E(y_0 | w)=E(y_0), prvi član sume u posljednjem retku prethodne jednadžbe iščezava. Također, ta pretpostavka implicira da odluka ljudi o tome žele li npr. sudjelovati u programu obuke za posao ne ovisi o tome koliko bi zarađivali ukoliko ne odu na taj program.

Kada bismo y_0 i y_1 zapisali kao: $y_0 = \mu_0 + \upsilon_0$ i $y_1 = \mu_1 + \upsilon_1$, gdje je μ_i= E(y_i), i=0, 1, dobili bismo:

y_1- y_0= (μ_1 - μ_0) + (υ_1- υ_0)=(E(y_1)-E(y_0))+ (υ_1- υ_0) = ATE + (υ_1- υ_0).

Djelujući na gornju jednadžbu s uvjetnim očekivanjem uz w =1, dobije se:

ATE_1= ATE + E(υ_1- υ_0 | w=1)

Na υ_1- υ_0 možemo gledati kao na korist od sudjelovanja, specifičnu za svaku osobu, pa se ATE i ATE_1 razlikuju za očekivanu korist onih koji su sudjelovali. Ako w i y_1 - y_0 nisu nezavisni, ATE i ATE_1 se općenito razlikuju.

1.3 Metode koje pretpostavljaju ”nezavisnost o tretmanu”

Neka **x** označava vektor opaženih kovarijabli. Na primjer, **x** može biti vektor koji nam govori koliko neko poduzeće ulaže u materijalnu/nematerijalnu imovinu, koliki je broj zaposlenih, da li postoji odjel istraživanja i razvoja i sl. Ako gledamo za osobu, $\mathbf{x}$ može sadržavati njezinu dob, stručnu spremu ili iskustvo.

Populaciju tada opisujemo sa $(y_0, y_1, w, \mathbf{x})$ i opažamo y, w i $\mathbf{x}$, gdje je y zadan sa (1.3). Kada nema uvjeta na koreliranost između w i (y_0, y_1), potrebne su dodatne pretpostavke kako bismo mogli identificirati efekt tretmana. Rosenbaum i Rubin (1983) (vidi [7]) predstavili su sljedeću pretpostavku i nazvali je **"nezavisnost o tretmanu"** *(eng. ignorability of treatment)* uz danu kovarijablu $\mathbf{x}$:

PRETPOSTAVKA ATE.1: Uvjetno na $\mathbf{x}$, w i (y_0, y_1) su nezavisni.

Također, možemo gledati i **"nezavisnost o tretmanu u smislu uvjetnog očekivanja"** (*eng. ignorability in a conditional mean independence sense*):

PRETPOSTAVKA ATE.1': (a) $E(y_0 \mid \mathbf{x}, w) = E(y_0 \mid \mathbf{x})$; (b) $E(y_1 \mid \mathbf{x}, w) = E(y_1 \mid \mathbf{x})$.

Pretpostavka ATE.1 povlači ATE.1'.

U pozadini, ideja je sljedeća: ako je moguće opaziti dovoljno informacija (koje su sadržane u $\mathbf{x}$) tako da je jedinstveno određeno hoće li pojedinac ući u tretman ili ne, onda možemo zaključiti da su, uvjetno na $\mathbf{x}$, w i (y_0, y_1) nezavisni.

Uzmimo za primjer studente koji se prijavljuju za obavljanje studentske stručne prakse. Neka vektor $\mathbf{x}$ sadržava informacije o njihovom prosjeku ocjena, dobi, godini studija te na kojem fakultetu studiraju. Ako pretpostavimo da $\mathbf{x}$ jedinstveno određuje tretman, onda bi trebalo vrijediti da 2 osobe sa istim vrijednostima $\mathbf{x}$ imaju isti ishod - ili su obje osobe primljene na praksu ili su obje odbijene.

ATE.1 sigurno vrijedi ako je w deterministička funkcija od $\mathbf{x}$, dakle ako je w određen sa $\mathbf{x}$ i samo sa $\mathbf{x}$, a sve komponente $\mathbf{x}$ su opazive. Ako je w=g($\mathbf{x}$, a), gdje je a neopaziva slučajna varijabla neovisna od $\mathbf{x}$ i ne utječe na y_0 i y_1, onda ATE.1 i dalje vrijedi. Vratimo se primjeru sa studentskom stručnom praksom: neka a označava koliki je interes studenta za obavljanje prakse, a y plaću osobe godinu dana nakon završetka studija, tj. y_0 plaću osobe koja tokom studija nije bila na praksi, a y_1 plaću osobe koja je. Neka je w=g($\mathbf{x}$, a). Ovisno o (gore navedenim) opazivim vrijednostima $\mathbf{x}$ i neopazivoj varijabli a, osoba će biti primljena na praksu ili neće. Ipak, a neće imati nikakvog utjecaja na (buduću) vrijednost y. Možemo zaključiti da su y i a nezavisni, te da ATE.1 vrijedi.

Važna činjenica je da pod pretpostavkom ATE.1' vrijedi jednakost prosječnog efekta tretmana uvjetno na $\mathbf{x}$ i prosječnog efekta tretmana tretiranih uvjetno na $\mathbf{x}$:

$$ATE_1(\mathbf{x}) \equiv E(\, y_1 - y_0 \mid \mathbf{x}, w = 1) = E(\, y_1 - y_0 \mid \mathbf{x}) = ATE(\mathbf{x}) \tag{1.6}$$

Ipak, jednakost ATE i ATE_1 (bez uvjeta na **x**) općenito ne vrijedi. Naime, ako definiramo r(**x**):= E(y_1 - y_0 | **x**) = ATE(**x**), ATE je očekivana vrijednost r(**x**) po čitavoj populaciji, dok je ATE_1 očekivana vrijednost r(**x**) po tretiranoj subpopulaciji. Matematički:

ATE = E[r(**x**)], ATE_1=E[r(**x**) | w=1].

Ako možemo procijeniti r(·), ATE se lako dobije usrednjavanjem po cijeloj populaciji, a ATE_1 po onom dijelu uzorka za koji je w=1.

1.3.1 Regresijske metode

Koristeći jednadžbu (1.3) i pretpostavku ATE.1' možemo dobiti procjenitelja za ATE(**x**) i zatim ga iskoristiti za procjenu ATE i ATE_1. Prvo,

E(y | **x**, w) = E((1 - w)y_0 + wy_1 | **x**, w) =

=E(y_0 | **x**, w) + w[E(y_1 | **x**, w) - E(y_0 | **x**, w)] = (zbog pretpostavke ATE.1')=

=E(y_0 | **x**) + w[E(y_1 | **x**) - E(y_0 | **x**)]

Također, iz ATE.1',

$$E(\, y \mid \mathbf{x}, w = 1) - E(\, y \mid \mathbf{x}, w = 0) = E(\, y_1 \mid \mathbf{x}) - E(\, y_0 \mid \mathbf{x}) = ATE(\mathbf{x}) \tag{1.7}$$

Zbog toga što je naš uzorak slučajan na (y, w, **x**), r_1(**x**) ≡ E(y | **x**, w=1) i r_0(**x**) ≡ E(y | **x**, w=0) su **neparametarski identificirane**, tj. ta dva uvjetna očekivanja ovise samo o opazivim vrijednostima varijabli i općenito se mogu konzistentno procijeniti. Stoga, možemo pretpostaviti da su r_1(**x**) i r_0(**x**) poznati, pa iz (1.7) vidimo da je i ATE(**x**) poznat.

Ako su $\hat{r}_1$(**x**) i $\hat{r}_0$(**x**) konzistentni procjenitelji i imamo slučajan uzorak veličine N, konzistentan procjenitelj za ATE je:

$$A\hat{T}E = N^{-1} \sum_{i=1}^{N} [\hat{r}_1(\mathbf{x}_i) - \hat{r}_0(\mathbf{x}_i)] \tag{1.8}$$

A konzistentan procjenitelj za ATE_1 je:

$$A\hat{T}E_1 = (\sum_{i=1}^{N} w_i)^{-1} \left\{ \sum_{i=1}^{N} w_i \, [\hat{r}_1(\mathbf{x}_i) - \hat{r}_0(\mathbf{x}_i)] \right\} \tag{1.9}$$

Formula (1.9) za $A\hat{T}E_1$ usrednjuje [$\hat{r}_1(\mathbf{x}_i)$ - $\hat{r}_0(\mathbf{x}_i)$] za poduzorak u kojem je w_i=1.

Postoji nekoliko problema kod računanja $A\hat{T}E$ i $A\hat{T}E_1$. Prvi je kako dobiti $\hat{r}_1(\cdot)$ i $\hat{r}_0(\cdot)$. Neke od varijanti su korištenje polinoma niskog stupnja koji uključuje i interakcije varijabli (npr. $x_1 \cdot x_2$) te korištenje logita i probita za binarne y. Ako pak imamo mnogo podataka, mogli bismo popisati sve moguće vrijednosti koje $\mathbf{x}$ može poprimiti - neka su to $\mathbf{c}_1$, $\mathbf{c}_2$, ..., $\mathbf{c}_M$ i procijeniti E(y | $\mathbf{x}$= $\mathbf{c}_m$, w=1) i E(y | $\mathbf{x}$= $\mathbf{c}_m$, w=0). Za svaki m i w=1 ili 0, ovom metodom procijenimo očekivanje preko srednje vrijednosti uzorka. Tipično, $\mathbf{x}$ poprima mnogo različitih vrijednosti pa je M velik.

Bez obzira na način dobivanja $\hat{r}_1(\cdot)$ i $\hat{r}_0(\cdot)$, za korištenje procjenjenog efekta tretmana bitno je da su dobivene greške asimptotski standardne normalne (*eng. asymptotically valid standard errors*).

To možemo postići korištenjem linearnog regresijskog modela, što ćemo pokazati kasnije u radu. Ipak, prije toga, bitno je uočiti probleme koji mogu nastati pri evaluaciji programa, posebice pri procjeni E(y | $\mathbf{x}$, w=1) i E(y | $\mathbf{x}$, w=0). Pretpostavit ćemo da postoji samo jedna binarna kovarijabla x, te da vrijedi pretpostavka ATE.1'. Na primjer, x može biti indikator koji pokazuje ako je dohodak pojedinca (prije tretmana) ispod određenog praga.

Također pretpostavimo da svatko sa x=1 sudjeluje u programu, tj. na svakog od njih je primjenjen tretman. Očito je da E(y | x=1, w=1) možemo procijeniti, a da E(y | x=1, w=0) ne možemo jer nemamo podataka za subpopulaciju sa $x = 1$ i $w = 0$. Preciznije, za članove populacije sa $x = 1$ opažamo y_1, no ne i y_0. Stoga, ATE(x) nije poznat za $x = 1$.

Ako za osobe za koje je $x = 0$ vrijedi da je na neke od njih primjenjen tretman, a na neke nije, ono što možemo procijeniti je E(y | x=0, w=1) - E(y | x=0, w=0), tj. možemo procijeniti ATE(x) za $x = 0$.

Dakle, budući da ne možemo procijeniti ATE(x) za $x = 1$, ne možemo procijeniti ni bezuvjetni ATE=P($x = 0$) · ATE(0) + P($x = 1$) · ATE(1).

Slične zaključke možemo dobiti i ako gledamo drugu situaciju u kojoj grupa sa $x = 0$ nije podvrgnuta tretmanu. Tada ne možemo procijeniti ATE(0) jer ni E(y | x=0, w=1) nije procjenjivo. Ako pak za grupu sa $x = 1$ imamo da su neki članovi povrgnuti tretmanu, a neki nisu, onda je ATE(1) poznat.

Ipak, postoji razlika između prve i druge situacije. Kao što smo već prije argumentirali, ako je dohodak neke osobe prije tretmana "velik", tj. vrijedi $x = 0$, vjerojatno nas neće zanimati podvrgavanje te osobe tretmanu obuke za posao. Preciznije, u situacijama kao u drugom slučaju, ono što nas zanima je zapravo samo ATE(1), a njega možemo procijeniti.

Možemo gledati i malo modificiranu situaciju. Neka je $\mathbf{x}$ binarni vektor koji označava u kojem je intervalu dohodak osobe prije tretmana. Pretpostavimo da je za osobe s najvećom razinom dohotka prije tretmana vjerojatnost sudjelovanja u tretmanu jednaka 0. To znači da možemo isključiti navedenu skupinu iz populacije. Također, ako je za osobe s najnižom razinom dohotka prije tretmana vjerojatnost sudjelovanja u tretmanu jednaka 1, i tu grupu ćemo isključiti iz populacije.

Općenito, gledajući slučajan uzorak, možemo odrediti je li vjerojatnost sudjelovanja u programu jednaka 1 ili 0. Ako su sve moguće vrijednosti koje $\mathbf{x}$ može poprimiti $\mathbf{c}_1$, $\mathbf{c}_2$, ..., $\mathbf{c}_M$, problem nastaje kada vrijedi da jedinke sa $\mathbf{x}= \mathbf{c}_m$ uvijek sudjeluju u programu. Iz razloga što ne možemo procijeniti E(y | $\mathbf{x}=\mathbf{c}_m$, w=0), subpopulaciju sa $\mathbf{x}= \mathbf{c}_m$ moramo isključiti iz analize.

Vratimo se sada na regresijske metode za procjenu ATE. Prikažimo y_0 i y_1 kao:

$$y_0 = \mu_0 + \upsilon_0, \quad E(\upsilon_0) = 0 \tag{1.10}$$

$$y_1 = \mu_1 + \upsilon_1, \quad E(\upsilon_1) = 0 \tag{1.11}$$

Iz prethodne dvije jednadžbe vidimo da je E(y_0) = μ_0 i E(y_1) = μ_1 te jednadžba (1.3) tada postaje:

$$y = (1 - w)y_0 + wy_1 = \mu_0 + (\mu_1 - \mu_0)w + \upsilon_0 + w(\upsilon_1 - \upsilon_0) \tag{1.12}$$

Ovo je primjer **"switching" regresijskog modela**: ovisno o statusu tretmana (tj. o vrijednosti w), y će poprimiti različite vrijednosti.

Propozicija 1.3.1. *Neka vrijedi ATE.1' i neka je E(* υ_1 | $\boldsymbol{x}$*) = E(* υ_0 | $\boldsymbol{x}$*). Tada je:* ATE_1=*ATE i E(* y | w, $\boldsymbol{x}$*) =* $\mu_0 + \alpha w + g_0(\boldsymbol{x})$ *gdje je* $\alpha \equiv$ *ATE i* $g_0(\boldsymbol{x})$ = *E(* υ_0 | $\boldsymbol{x}$*).*

Ako dodatno vrijedi da je E(υ_0 | $\boldsymbol{x}$*) =* $\eta_0 + \boldsymbol{h}_0(\boldsymbol{x})\beta_0$ *za vektorsku funkciju* $\boldsymbol{h}_0(\boldsymbol{x})$*, onda E(* y | w, $\boldsymbol{x}$*) =* $\gamma_0 + \alpha w + \boldsymbol{h}_0(\boldsymbol{x})\beta_0$*, gdje je* $\gamma_0 = \mu_0 + \eta_0$.

Dokaz. Zbog pretpostavke ATE.1' vrijedi:
E(y_i | w, $\mathbf{x}$) = E(y_i | $\mathbf{x}$) = E($\mu_i + \upsilon_i$ | $\mathbf{x}$) = μ_i+ E(υ_i | $\mathbf{x}$), i = 0, 1.

Iz pretpostavke E(υ_1 | **x**) = E(υ_0 | **x**) slijedi:
E(y_1 | w, **x**) - E(y_0 | w, **x**) = μ_1 + E(υ_1 | **x**) - μ_0 - E(υ_0 | **x**) = μ_1 - μ_0.

Dakle, imamo da je ATE = E(y_1 - y_0) = μ_1 - μ_0, ATE_1= E(y_1 - y_0 | w)= (LIE) = E(E(y_1 - y_0 | w, **x**) | w) = μ_1 - μ_0, pa zaključujemo da je ATE = ATE_1.

Za dokazivanje druge tvrdnje pustimo uvjetno očekivanje, uz dane **x** i w na jednadžbu (1.12), uz korištenje pretpostavki ATE.1' i E(υ_1 | **x**) = E(υ_0 | **x**):

$$\begin{aligned} E(\, y \mid w,\ \mathbf{x}) &= E(\, \mu_0 + (\mu_1 - \mu_0)w + \upsilon_0 + w(\upsilon_1 - \upsilon_0) \mid w,\ \mathbf{x}) \\ &= \mu_0 + (\mu_1 - \mu_0)w + E(\, \upsilon_0 \mid w,\ \mathbf{x}) + wE(\upsilon_1 - \upsilon_0 \mid w,\ \mathbf{x}) = (ATE.1') \\ &= \mu_0 + (\mu_1 - \mu_0)w + E(\, \upsilon_0 \mid\ \mathbf{x}) + wE(\upsilon_1 - \upsilon_0 \mid\ \mathbf{x}) \\ &= \mu_0 + (\mu_1 - \mu_0)w + E(\, \upsilon_0 \mid\ \mathbf{x}) + 0 \end{aligned}$$

Dakle, pokazali smo relaciju E(y | w, **x**) = μ_0 + αw + g_0(**x**) za $\alpha \equiv$ ATE i
g_0(**x**) = E(υ_0 | **x**).

Posljednja tvrdnja je samo specijalni slučaj prethodne. □

Prethodna propozicija nam govori da je, kada vrijedi E(υ_1-υ_0 | **x**) = 0, E(y | w, **x**) aditivno po w i funkcija od **x**, a koeficijent uz w je prosječan efekt tretmana. Stoga možemo koristiti regresijske metode za procjenu ATE. Ukoliko je E(υ_0 | **x**) nelinearna u parametrima, mogu se koristiti nelinearne regresijske metode. Ipak, tipično je
E(υ_0 | **x**) oblika E(υ_0 | **x**) = η_0 + $\mathbf{h}_0(\mathbf{x})\beta_0$ za vektorsku funkciju $\mathbf{h}_0(\mathbf{x})$, kao u prethodnoj propoziciji. Korištenje fleksibilnih funkcijskih oblika kao elemenata u $\mathbf{h}_0(\mathbf{x})$ trebalo bi dati dobru aproksimaciju za E(υ_0 | **x**).

Problem je što E(υ_1-υ_0 | **x**) = 0 općenito ne vrijedi. Vrijedit će, na primjer, u situacijama u kojima je $\upsilon_1 = \upsilon_0$ ili $y_1 = y_0 + \alpha$, za $\alpha \equiv$ ATE, što znači da je efekt tretmana jednak za sve članove populacije. No, i bez tog uvjeta moguće je procijeniti ATE:

Propozicija 1.3.2. *Neka vrijedi ATE.1'. Tada je:*
*E(y | w, **x**) = μ_0 + αw + g_0(**x**) + w[g_1(**x**) - g_0(**x**)] gdje je $\alpha \equiv$ ATE,*
*g_0(**x**) $\equiv$ E(υ_0 | **x**) i g_1(**x**) $\equiv$ E(υ_1 | **x**).*

Dokaz. Djelovanjem uvjetnim očekivanjem uz dane **x** i w na jednadžbu (1.12) dobijemo:

$$\begin{aligned} E(\, y \mid w, \mathbf{x}) &= E(\, \mu_0 + (\mu_1 - \mu_0)w + \upsilon_0 + w(\upsilon_1 - \upsilon_0) \mid w, \mathbf{x}) \\ &= \mu_0 + (\mu_1 - \mu_0)w + E(\, \upsilon_0 \mid w,\ \mathbf{x}) + wE(\upsilon_1 - \upsilon_0 \mid w,\ \mathbf{x}) = (ATE.1') \\ &= \mu_0 + (\mu_1 - \mu_0)w + E(\, \upsilon_0 \mid\ \mathbf{x}) + wE(\upsilon_1 - \upsilon_0 \mid\ \mathbf{x}) \end{aligned}$$

Time smo dokazali navedenu tvrdnju. □

Iz jednadžbe E(y | w, **x**) = μ_0 + αw + $g_0(\mathbf{x})$ + w[$g_1(\, \mathbf{x})$ - $g_0(\, \mathbf{x})$] vidimo da je E(y | w, **x**) aditivna u w, funkcija od **x** i interakcija w i **x**. Koeficijent uz w predstavlja ATE. Ako su υ_1 i υ_0 nepoznati, u parametarskom okruženju, $g_0(\cdot)$ i $g_1(\cdot)$ zamijenit ćemo parametarskim funkcijama **x**, tipično linearnima u parametrima, npr. η_0 + $\mathbf{h}_0(\mathbf{x})\beta_0$ i η_1 + $\mathbf{h}_1(\mathbf{x})\beta_0$. Ako pretpostavimo da su obje linearne u **x**, možemo napisati:

E(y | w, **x**) = γ + αw + $\mathbf{x}\boldsymbol{\beta}_0$ + w(**x**- $\boldsymbol{\psi}$)$\boldsymbol{\delta}$

gdje su $\boldsymbol{\beta}_0$ i $\boldsymbol{\delta}$ vektori nepoznatih parametara i $\boldsymbol{\psi} \equiv$ E(**x**). Oduzimanjem očekivanja od **x**, osigurali smo da je ATE koeficijent uz w. U praksi, od **x** možemo oduzeti srednju vrijednost populacije (ako nam je poznata) ili srednju vrijednost uzorka. α tada procijenimo iz regresije

$$y_i \quad na \quad 1, w_i, \mathbf{x}_i, w_i(\mathbf{x}_i - \bar{\mathbf{x}}), \quad i = 1, 2, ..., N \tag{1.13}$$

gdje je $\bar{\mathbf{x}}$ vektor sa srednjim vrijednostima uzorka. Kontrolne funkcije, u ovom slučaju, ne uključuju samo $\mathbf{x}_i$ nego i interakcije $\mathbf{x}_i$ sa w.

Zbog toga što regresija (1.13) konzistentno procjenjuje $\boldsymbol{\delta}$, možemo gledati kako se ATE(**x**) = E(y_1 - y_0 | **x**) mijenja s elementima **x**. Točnije, koeficijent uz w te regresije procjenjuje ATE pa je:

$A\hat{T}E(\mathbf{x}) = \hat{\alpha} + (\, \mathbf{x} \text{-}\ \bar{\mathbf{x}}\,)\hat{\boldsymbol{\delta}}$

Gledajući po raznim **x**, možemo procijeniti ATE za razne podskupove populacije. Ako **x** sadrži informaciju o dohotku prije tretmana ili je **x** indikator kojoj grupi (u odnosu na razinu dohotka) osoba pripada, možemo procijeniti kako se ATE mijenja za razne razine dohotka prije tretmana. Problem je ako su odabrane grupe "premale". Naime, ukoliko su grupe složene tako da se u prvoj grupi nalaze ljudi s dohotkom od 2000 - 3000 kn, u idućoj od 3000 - 4000 kn i tako redom, teže će biti međusobno uspoređivati ATE(**x**) tih grupa, nego kad bismo imali veće grupe u kojima je raspon dohotka 5000 kn.

Za procijeniti ATE_1= E(y_1 - y_0 | w=1), napišemo E(y | w, **x**) kao:

E(y | w, **x**) = (*LIE*) = E(E(y | w, **x**) | w).

Uz korištenje E(y | w, **x**) = γ + αw + $\mathbf{x}\boldsymbol{\beta}_0$ + w(**x**- $\boldsymbol{\psi}$)$\boldsymbol{\delta}$, pokaže se da je

$ATE_1 = \alpha$ + [E(**x** | w=1) - $\boldsymbol{\psi}$]$\boldsymbol{\delta}$ pa je konzistentan procjenitelj:

$$A\hat{T}E_1 = \hat{\alpha} + (\sum_{i=1}^{N} w_i)^{-1} [\sum_{i=1}^{N} w_i (\mathbf{x}_i - \bar{\mathbf{x}})\hat{\boldsymbol{\delta}})] \tag{1.14}$$

Problem je što je procjena standardnih grešaka dosta komplicirana, no može se izvesti korištenjem delta metode ili "bootstrapping"-a.

Primjer 1.3.3. *(Utjecaj poduzetničkih zona na ekonomski razvoj)*

Poduzetničke zone (*eng. enterprise zone* - EZ) kreiraju se kako bi se povećala zaposlenost i tokovi kapitala u ekonomski slabijim područjima države. To su područja s velikom nezaposlenosti, niskim razinama dohotka, brojnim neiskorištenim površinama i zastarjelom infrastrukturom. Investitorima se daju porezne olakšice ili subvencije, reduciraju se regulative i potrebna administracija i sl.

Pretpostavimo da promatramo učinak tih zona na rast zaposlenosti *(gemp = employment growth)* i da je kreiranje zona započelo početkom 1980-ih, te da imamo popis podataka za 1980-e i 1990-e. Vodeći računa o tome da označavanje zona vjerojatno ovisi o prethodnim ekonomskim učincima, kao i o obilježjima područja u kojem se nalazi, možemo procijeniti model poput sljedećeg:

$$\begin{aligned} gemp = {} & \mu_0 + \alpha ez + \beta_1 log(emp80) + \beta_2 log(pop80) + \beta_3 percmanf80 \\ & + \beta_4 log(housval80) + \beta_5 ez \cdot [log(pop80) - m_1] + \beta_6 ez \cdot [log(pop80) - m_2] \\ & + \beta_7 ez \cdot [percmanf80 - m_3] + \beta_8 ez \cdot [log(housval80) - m_4] + error \end{aligned} \tag{1.15}$$

Varijable na desnoj strani jednadžbe su dummy varijable koje predstavljaju redom: označavanje EZ zone, zaposlenost, populaciju, stopu zaposlenih u proizvodnji, prosječnu vrijednost nekretnina, sve u 1980-ima, te m_j predstavljaju srednje vrijednosti uzoraka.

U ovom slučaju, tretman je određen nestohastičkom funkcijom kovarijable, npr. $w = f(s)$, gdje je s element **x** sa zadovoljavajućom varijacijom. Ključno je da je f diskretna funkcija s, tipično step funkcija w= 1[s $\leq s_0$] gdje je s_0 neki poznati prag, a s razina dohotka ili veličina razreda. Ideja je da čim s prijeđe s_0, program automatski započinje. Zbog toga što je s neslučajna funkcija **x**, a w je određen sa s, zaključujemo da su w i (y_0, y_1) nezavisni uvjetno na **x**, tj. pretpostavka ATE.1 vrijedi. Ključno je pretpostaviti da su $g_0(\cdot)$ i $g_1(\cdot)$ glatke funkcije **x** kako bismo mogli identificirati α tj. kako bismo uopće mogli primijeniti regresiju (1.13)

1.3.2 Metode bazirane na vjerojatnosti sklonosti

Definicija 1.3.4. *Funkciju p(**x**) definiranu sa p(**x**) ≡ P(w = 1 | **x**) nazivamo **propensity score** ili **vjerojatnost sklonosti**.*

$p(\mathbf{x})$ je dakle vjerojatnost da jedinka primi tretman ako nam je poznat vektor kovarijabli **x**. Iz sljedeće propozicije vidimo da se ATE i ATE_1 mogu prikazati u terminima vjerojatnosti sklonosti:

Propozicija 1.3.5. *Neka vrijedi pretpostavka ATE.1' i neka je 0 < p(**x**) < 1, ∀ **x**. Tada je:*
$ATE = E([\, w - p(\mathbf{x})\,]\, y / \{ p(\mathbf{x})\, [1 - p(\mathbf{x})\,] \})$
$ATE_1 = E \{ [\, w - p(\mathbf{x})\,]\, y / [\, 1 - p(\mathbf{x})\,] \} / P(w = 1)$

Dokaz. U izrazu [w - $p(\mathbf{x})$] y supstituiramo y sa y=(1-w)y_0 + wy_1 te dobijemo:

$$\begin{aligned}[w - p(\mathbf{x})\,]y &= [w - p(\mathbf{x})\,][\, (1 - w)y_0 + wy_1] \\ &= wy_1 - p(\mathbf{x})(1 - w)y_0 - p(\mathbf{x})wy_1\end{aligned}$$

Na gornju jednadžbu djelujemo uvjetnim očekivanjem, uvjetno na (w, **x**) i koristeći pretpostavku ATE.1' dobijemo:

$$wE(\, y_1 \mid w, \mathbf{x}) - p(\mathbf{x})(1 - w)E(\, y_0 \mid w, \mathbf{x}) - p(\mathbf{x})wE(\, y_1 \mid w, \mathbf{x})$$

Ako označimo $m_j(\mathbf{x}) \equiv \mathrm{E}(\, y_j \mid \mathbf{x})$, j = 0, 1, slijedi:

$$wm_1(\mathbf{x}) - p(\mathbf{x})(1 - w)m_0(\mathbf{x}) - p(\mathbf{x})wm_1(\mathbf{x})$$

Gledajući uvjetno očekivanje, uz dani **x**, iz gornje relacije, koristeći da je $\mathrm{E}(\, w \mid \mathbf{x}) = p(\mathbf{x})$ slijedi:

$$p(\mathbf{x})m_1(\mathbf{x}) - p(\mathbf{x})[1 - p(\mathbf{x})]m_0(\mathbf{x}) - [p(\mathbf{x})]^2 m_1(\mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})][m_1(\mathbf{x}) - m_0(\mathbf{x})]$$

Dakle, pokazali smo da vrijedi:
E([w - $p(\mathbf{x})$] y / { $p(\mathbf{x})$ [1- $p(\mathbf{x})$] } | $\mathbf{x}$) = (LIE) =
= E(E([w - $p(\mathbf{x})$] y / { $p(\mathbf{x})$ [1- $p(\mathbf{x})$] } | $\mathbf{x}$,w) | $\mathbf{x}$) = [$m_1(\mathbf{x})$- $m_0(\mathbf{x})$]
= E(y_1 - y_0 | $\mathbf{x}$) = μ_1 - μ_0 = ATE

pa prva tvrdnja propozicije vrijedi.

Na sličan način pokaže se da je:
E { [w - $p(\mathbf{x})$] y / [1- $p(\mathbf{x})$] | $\mathbf{x}$ } = $p(\mathbf{x})[m_1(\mathbf{x})$- $m_0(\mathbf{x})]$

Iteriranjem očekivanja slijedi:
E($p(\mathbf{x})$ [$m_1(\mathbf{x})$- $m_0(\mathbf{x})$]) = E(E(w | $\mathbf{x}$) E(y_1 - y_0 | $\mathbf{x}$)) = (zbog ATE.1') = E(w (y_1 - y_0))

Međutim,

$$\begin{aligned} E(w(y_1 - y_0)) &= P(w = 1)E[w(y_1 - y_0)|w = 1] + P(w = 0)E[w(y_1 - y_0)|w = 0] \\ &= P(w = 1)E[w(y_1 - y_0)|w = 1] \end{aligned}$$

Stoga je :
E($p(\mathbf{x})$ [$m_1(\mathbf{x})$- $m_0(\mathbf{x})$]) = E (y_1 - y_0 | w = 1) = ATE_1,
čime smo dokazali i drugu tvrdnju. □

Pretpostavku ATE.1 zajedno s uvjetom $0 < p(\mathbf{x}) < 1$, $\forall$ $\mathbf{x}$, Rosenbaum i Rubin (1983) (vidi [7]) nazivaju **jaka nezavisnost o tretmanu** (*eng. strong ignorability of a treatment*) (uz danu kovarijablu $\mathbf{x}$).

Iz prethodne propozicije vidimo da su ATE i $ATE.1$ neparameterski identificirani ukoliko vrijedi jaka nezavisnost o tretmanu. Isključivanje mogućnosti da $p(\mathbf{x}) = 0$ i $p(\mathbf{x}) = 1$ je logično jer ionako želimo isključiti jedinke koje nikad neće biti tretirane, kao i one koje su uvijek tretirane.

Da bismo procijenili ATE i ATE_1, prvo trebamo procijeniti $p(\cdot)$. Rosenbaum i Rubin (1983) sugeriraju korištenje fleksibilnog logit modela u kojem uključimo razne funkcije $\mathbf{x}$-a, kao što su kvadrati ili interakcije. U tom slučaju, budući da se radi o logit modelu, sigurno nećemo dobiti da je $\hat{p}(\mathbf{x}) = 0$ ili 1. Vjerojatnost sklonosti također možemo procijeniti i neparametarskim metodama. Ako je procjenitelj oblika $\hat{p}(\mathbf{x}) \equiv F(\ \mathbf{x}; \hat{\boldsymbol{\gamma}}\ )$, gdje je $\hat{\boldsymbol{\gamma}}$ dobiven iz binarnog modela w, na $\mathbf{x}$, konzistentan procjenitelj za ATE je:

$$A\hat{T}E = N^{-1} \sum_{i=1}^{N} [w_i - \hat{p}(\mathbf{x}_i)]\ y_i\ /\ \{\ \hat{p}(\mathbf{x}_i)\ [1 - \hat{p}(\mathbf{x}_i)]\ \} \tag{1.16}$$

Budući da je ($N^{-1}\sum_{i=1}^{N} w_i$) konzistentan procjenitelj za P($w = 1$), konzistentan procjenitelj za ATE_1 je:

$$A\hat{T}E_1 = (N^{-1} \sum_{i=1}^{N} w_i)^{-1} \left\{ N^{-1} \sum_{i=1}^{N} [w_i - \hat{p}(\mathbf{x}_i)]\ y_i\ /\ [1 - \hat{p}(\mathbf{x}_i)] \right\} \tag{1.17}$$

Iz jednadžbi (1.16) i (1.17) vidimo da ukoliko dvije metode imaju slične predviđene vrijednosti $\hat{p}(\mathbf{x}_i)$, i procjene za efekt tretmana će biti slične. Procjenitelji dani tim dvjema jednadžbama imat će najmanju asimptotsku varijancu od svih procjenitelja baziranih samo na pretpostavci ATE.1 i uvjetu $0 < p(\mathbf{x}) < 1,\ \forall\ \mathbf{x}$.

Za procjenu efekta tretmana možemo koristiti i običnu metodu najmanjih kvadrata (OLS) u kojoj radimo regresiju:

$$y_i\ na\ 1, w_i, \hat{p}(\mathbf{x}_i), \quad i = 1, 2, ..., N \tag{1.18}$$

Koeficijent uz w_i tada predstavlja prosječan efekt tretmana, dok $\hat{p}(\mathbf{x}_i)$ ima ulogu kontrolne funkcije. Ideja je da vjerojatnost sklonosti sadržava sve one informacije iz kovarijabli koje su značajne za procjenu prosječnog efekta tretmana.

Sljedeća propozicija (koju navodimo bez dokaza) govori kada takva regresija konzistentno procjenjuje efekt tretmana.

Propozicija 1.3.6. *Neka vrijedi ATE.1' i neka je E(y_1 - y_0 | $\boldsymbol{x}$) = $m_1(\boldsymbol{x})$ - $m_0(\boldsymbol{x})$ nekorelirano sa Var(w | $\boldsymbol{x}$) = $p(\boldsymbol{x})$ [1- $p(\boldsymbol{x})$]. Ako je procjenitelj $\hat{p}(\cdot)$ konzistentan i $\sqrt{N}$- asimptotski normalan, tada je koeficijent uz w regresije:*

y_i na 1, w_i, $\hat{p}(\boldsymbol{x}_i)$, i= 1, 2, ..., N

konzistentan i $\sqrt{N}$- asimptotski normalan za ATE.

Pretpostavka da su $m_1(\mathbf{x})$ - $m_0(\mathbf{x})$ i Var(w | $\mathbf{x}$) nekorelirani na prvu se može činiti nemoguća, budući da su oba izraza funkcije od $\mathbf{x}$. Ipak, moramo se sjetiti da je korelacija linearna mjera zavisnosti. Budući da je Var(w | $\mathbf{x}$) = p($\mathbf{x}$) [1- p($\mathbf{x}$)] nemonotona i kvadratna u $\hat{p}(\mathbf{x})$, dok je $m_1(\mathbf{x})$ - $m_0(\mathbf{x})$ vjerojatno monotona u mnogim elementima $\mathbf{x}$, aproksimativno bi mogla vrijediti nekoreliranost (to je analogno činjenici da ako je z standardna normalna slučajna varijabla, da su z i z^2 nekorelirane).

Rosenbaum i Rubin (vidi [7], korolar 4.3) predlažu korištenje malo općenitije varijante regresije kako bismo procijenili ATE. Ta regresija je oblika:

$$y_i \; na \; 1, w_i, \hat{p}_i, w_i(\; \hat{p}_i - \hat{\mu}_p \;) \quad i = 1, 2, ..., N \tag{1.19}$$

$\hat{\mu}_p$ je srednja vrijednost uzorka od $\hat{p}_i$ za i= 1, 2, ..., N

Propozicija 1.3.7. *Neka vrijedi ATE.1 i neka su E[y_0 | p($\mathbf{x}$)] i E[y_1 | p($\mathbf{x}$)] linearni u p($\mathbf{x}$). Tada koeficijent uz w_i u regresiji* (1.19) *konzistentno procjenjuje ATE.*

Dokaz. Rosenbaum i Rubin (vidi [7], Teorem 3) pokazali su da su pod pretpostavkom ATE.1, (y_0, y_1) i w nezavisni uvjetno na $p(\mathbf{x})$. Da bi to vrijedilo, dovoljno je da vrijedi P[w=1 | y_0, y_1, $p(\mathbf{x})$] = P[w=1 | $p(\mathbf{x})$] ili E[w | y_0, y_1, $p(\mathbf{x})$] = E[w | $p(\mathbf{x})$]. Iz pretpostavke ATE.1 slijedi da je E[w | y_0, y_1, $\mathbf{x}$] = E[w | $\mathbf{x}$] = $p(\mathbf{x})$. Iz zakona iteriranih očekivanja slijedi:

E[w | y_0, y_1, $p(\mathbf{x})$] = (LIE) = E[E(w | $y_0, y_1, \mathbf{x}$) | $y_0, y_1, p(\mathbf{x})$] =
= E[$p(\mathbf{x})$ | $y_0, y_1, p(\mathbf{x})$] = $p(\mathbf{x})$.

Time smo pokazali da su uvjetno na $p(\mathbf{x})$, w i (y_0, y_1) nezavisni.

Ako napišemo y kao y= y_0 + (μ_1 - μ_0)w + (υ_1 - υ_0)w, dobivamo:

E[y | w, $p(\mathbf{x})$] = (jer su w i y_0 nezavisni uz dani $p(\mathbf{x})$) =
= E[y_0 | $p(\mathbf{x})$] + (μ_1 - μ_0)w + w(E[υ_1 | $p(\mathbf{x})$] - E[υ_0 | $p(\mathbf{x})$]) =
= (jer je E[y_0 | $p(\mathbf{x})$] linearna u $p(\mathbf{x})$) =
= $\delta_0 + \delta_1 p(\mathbf{x})$ + (μ_1 - μ_0)w + $\delta_2 w$[$p(\mathbf{x})$ - μ_p].

Zbog toga što su očekivanja od υ_1 i υ_0 jednaka 0, a E[y_1 | $p(\mathbf{x})$] i E[y_0 | $p(\mathbf{x})$] su linearni u $p(\mathbf{x})$, i E [υ_1 | $p(\mathbf{x})$] i E[υ_0 | $p(\mathbf{x})$] moraju imati očekivanja jednaka 0. Stoga posljednji izraz gornjoj jednadžbi možemo napisati kao $\delta_2 w$[$p(\mathbf{x})$ - μ_p].

Kada μ_p zamijenimo sa $\hat{\mu_p}$, tj. sa srednjom vrijednosti uzorka od $\hat{p}_i$, za i= 1, 2, ..., N, to ne utječe na konzinstentnost (ni na asimptotsku normalnost).

□

Primjer 1.3.8. *(Utjecaj obuke za posao na dohodak)*

Tokom 1970-ih, provođene su obuke za posao i gledan je njihov utjecaj na realni dohodak (y), mjeren u tisućama dolara, 1978. godine. Za osobe koje nisu radile tokom godine, $y = 0$. Za procjenu prosječnog efekta tretmana korištene su regresije oblika (1.18) i (1.19). Vektor $\mathbf{x}$ sastoji se od sljedećih elemenata: realni dohodak pojedinca u 1974. i 1975. godini, dob pojedinca u kvadratnom obliku, binarna varijabla koja označava ima li osoba završenu srednju školu (*nodegree*), binarna varijabla koja označava je li osoba u braku, te binarne varijable koje označavaju je li osoba crne rase ili Hispanac. Varijabla w označava je li osoba podvrgnuta tretmanu (obuci) ili ne.

U prvom koraku koristimo probit model za procjenu $\hat{p}(\mathbf{x}_i)$. Tretman je dakle zavisna varijabla, a elementi $\mathbf{x}$ eksplanatorne. Na razini značajnosti 5% jedino je varijabla *nodegree* statistički značajna. Zatim radimo regresiju (1.18) iz koje dobivamo da je $\hat{\alpha}$ = 1.626, sa standardnom greškom (koja nije prilagođena za procjenu iz probit-a) se = .644. Ono što zaključujemo je da dodatnom obukom realni dohodak pojedinca raste za otprilike 1626$.

Ako bismo radili regresiju oblika y na 1, w, $\mathbf{x}$, dobili bismo da je $\hat{\alpha}$ = 1.625, sa standardnom greškom se = .640, tj. dobiveni rezultati iz obje regresije su jako slični.

Kada bismo samo gledali razliku srednjih vrijednosti dohodaka nakon i prije obuke, dobili bismo da je razlika 1.794, sa standardnom greškom se = .633.

Iz regresije (1.19) sa $\hat{\mu}_p$ = .416, dobije se nešto niža vrijednost procjene: $\hat{\alpha}$ = 1.560, sa standardnom greškom se = .642, te vrijedi da je interakcijski dio w_i ($\hat{p}_i$ - $\hat{\mu}_p$) značajan.

Regresije (1.18) i (1.19) su "privlačne" jer se mogu koristiti i u situacijama kada tretman nije slučajno raspoređen po jedinkama, budući da uključuju $\hat{p}(\mathbf{x}_i)$ (tj. funkciju kovarijabli) kao samostalnu varijablu. U usporedbi s regresijama koje uključuju cijeli skup elemenata iz $\mathbf{x}$, eventualno u interakciji s indikatorom tretmana w (kao, na primjer, u regresiji (1.13)), mogu se ipak činiti pomalo "škrte". U stvarnosti, ukoliko vjerojatnost sklonosti procijenimo korištenjem linearnog vjerojatnosnog modela, regresije (1.18) i (1.13) bez interakcije varijabli dale bi iste procjene za α. Dok smo u prethodnom primjeru standardnu grešku procijenili ignorirajući početnu procjenu vjerojatnosti sklonosti, kod regresije (1.13) sigurni smo da je procjena standardne greške od $\hat{\alpha}$ pouzdana.

U prethodnom primjeru nismo mogli procijeniti utjecaj uzoračke varijacije, u procjeni iz probit modela, na standardnu grešku procjene prosječnog efekta tretmana. Kod modela koji koriste vjerojatnost sklonosti naspram standardnih regresijskih modela ne možemo pronaći dominaciju, budući da te dvije vrste modela zahtijevaju drukčije pretpostavke za konzistentnost. Nažalost, malo je toga napravljeno u području uspoređivanja regresijskih metoda koje koriste vjerojatnost sklonosti i standardnih **"kitchen sink"** tipova regresija, kao ni u usporedbi tih metoda s procjeniteljem danim jednadžbom (1.16).

Rosenbaum i Rubin (vidi [7]) predložili su pristup "uparivanja", motiviran sljedećom idejom: pretpostavimo da je vjerojatnost sklonosti $p(\mathbf{x})$ odabrana slučajnim odabirom iz populacije. Zatim odaberemo dvije jedinke koje imaju istu vrijednost $p(\mathbf{x})$, od kojih je jedna primila tretman, a druga nije. Pod pretpostavkom ATE.1, očekivana razlika u opaženim ishodima tih dvaju jedinki je:

$$E[\ y \mid w = 1, p(\mathbf{x})\ ] - E[\ y \mid w = 0, p(\mathbf{x})\ ] = (\ ATE.1\ ) = E[\ y_1 - y_0 \mid p(\mathbf{x})\ ]$$

Iteriranjem očekivanja dobijemo ATE = E(y_1 - y_0). Da bismo dakle procijenili ATE, potrebno je prvo procijeniti vjerojatnosti sklonosti, zatim upariti jedinke na temelju vjerojatnosti sklonosti i procijeniti razliku u ishodima između uparenih jedinki, te na kraju uprosječiti vrijednosti razlike ishoda svih parova. Budući da je teško pronaći dvije jedinke s istom vrijednosti $p(\mathbf{x})$, koristi se grupiranje ili lokalno uprosječivanje. To znači da se uparuju jedinke i sa sličnim vrijednostima $p(\mathbf{x})$. U praksi je ipak problem pronaći jedinke koje imaju slične procijenjene vjerojatnosti sklonosti, a da je jedna od njih tretirana, a druga nije.

1.4 Metode instrumentalnih varijabli

1.4.1 Motivacija za korištenje instrumentalnih varijabli

Pretpostavimo da promatramo sljedeći linearni model:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_K + u \tag{1.20}$$

E(u) = 0, Cov(x_j, u) = 0, j = 1, 2, ... , K-1, dok x_K i u mogu biti korelirani.

Eksplanatorne varijable x_1, x_2, ... , x_{K-1} su egzogene, no x_K je potencijalno endogena u gornjoj jednadžbi. O tome bismo mogli razmišljati na način da u sadrži neku izostavljenu varijablu za koju nismo imali dovoljno podataka, pa je nismo mogli uključiti u regresiju i koja je nekorelirana sa svim eksplanatornim varijablama, osim sa x_K. Zbog te koreliranosti, OLS metodom ne možemo konzistentno procijeniti koeficijente β_j. Metoda instrumentalnih varijabli (IV) rješava problem endogene eksplanatorne varijable x_K. Prvo, potrebno je naći varijablu z_1 koja se ne pojavljuje u jednadžbi (1.20) i koja zadovoljava sljedeća dva uvjeta:

$$Cov(z_1, u) = 0, \tag{1.21}$$

tj. z_1 je nekorelirana sa u, te

$$x_K = \delta_0 + \delta_1 x_1 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_K \tag{1.22}$$

gdje je $\mathrm{E}(r_K) = 0$, r_K je nekorelirana sa x_1, x_2, ... , x_{K-1}, z_1 i koeficijent $\theta_1 \neq 0$. Gornju jednadžbu nazivamo **jednadžba za x_K u reduciranoj formi**.

Varijabla z_1, za koju vrijede gore navedena dva uvjeta, naziva se **instrumentalna varijabla** za x_K ili **instrument** za x_K. Ukoliko je neka varijabla egzogena, ona je instrumentalna varijabla za "samu sebe", no pod pojmom instrumentalnih varijabli uglavnom se navode samo instrumenti za endogene eksplanatorne varijable.

U jednadžbi (1.22), endogenu varijablu x_K prikazali smo kao linearnu projekciju na egzogene varijable. Sada tu relaciju uvrštavamo u jednadžbu (1.20):

$$y = \alpha_0 + \alpha_1 x_1 + ... + \alpha_{K-1} x_{K-1} + \lambda_1 z_1 + v \tag{1.23}$$

gdje je $v = u + \beta_K r_K$, $\alpha_j = \beta_j + \beta_K \delta_j$, $\lambda_1 = \beta_K \theta_1$. Budući je, po konstrukciji, v nekoreliran sa svim eksplanatornim varijablama, OLS konzistentno procjenjuje koeficijente α_j i λ_1.

Metodu instrumentalnih varijabli možemo, na primjer, primijeniti u situaciji u kojoj promatramo učinak dodatne obuke radnika na produktivnost rada. Neka je x_K varijabla koja označava broj sati obuke po radniku, a y prosječnu produktivnost radnika. Pretpostavimo da se na slučajan način odabiru kompanije koje će primiti subvencije za dodatnu obuku radnika. Za varijablu z_1 mogli bismo tada uzeti binarnu varijablu koja označava je li kompanija primila subvenciju ili koliki je iznos subvencije po radniku. Parametar β_K iz jednadžbe (1.20) govori koliki je utjecaj obuke radnika na produktivnost rada, no njega ne možemo procijeniti. Ako je z_1 binarna varijabla i označava primanje subvencije, onda koeficijent λ_1 iz jednadžbe (1.23) pokazuje koliki je utjecaj primanja subvencije na produktivnost rada. Ipak, više bi nam koristila informacija o utjecaju sata obuke, nego o primanju subvencije. Stoga ćemo sada pokazati kako je moguće dobiti i koeficijente β_j.

Jednadžbu (1.20) zapisat ćemo kao:

$$y = \mathbf{x}\boldsymbol{\beta} + u \tag{1.24}$$

gdje je $\mathbf{x}$ = (1, x_1, ... , x_K). Označimo sa $\mathbf{z}$ = (1, x_1, ... , x_{K-1} , z_1) vektor svih egzogenih varijabli. Iz uvjeta (1.21) i (1.22) slijedi:

$$E(\mathbf{z}'u) = \mathbf{0} \tag{1.25}$$

Množenjem jednadžbe (1.24) sa **z**' i uzimanjem očekivanja dobivamo:

$$E(\mathbf{z}'\mathbf{x})\boldsymbol{\beta} = E(\mathbf{z}'y) \tag{1.26}$$

Ovaj sustav ima jedinstveno rješenje $\boldsymbol{\beta}$ ako je matrica E(**z**'**x**) punog ranga. U tom slučaju, rješenje je:

$$\boldsymbol{\beta} = E(\mathbf{z}'\mathbf{x})^{-1}E(\mathbf{z}'y) \tag{1.27}$$

Očekivanja E(**z**'**x**) i E(**z**'y) možemo konzistentno procijeniti uzimanjem slučajnog uzorka za (**x**, y, z_1).

1.4.2 2SLS metoda kod višestrukih instrumenata

Promatrajmo model kao u (1.20) i pretpostavimo da postoji više instrumentalnih varijabli za x_K i neka su to varijable z_1, z_2, ..., z_M. Za h= 1, 2, ..., M vrijedi:

$$Cov(z_h, u) = 0, \tag{1.28}$$

tj. z_h su egzogene u jednadžbi (1.20). Budući da su i linearne kombinacije od x_1, x_2, ..., x_{K-1}, z_1, z_2, ..., z_M nekorelirane sa u, potencijalnih instrumenata za x_K ima i više od M. Postavlja se pitanje kojeg od svih mogućih instrumenata je najbolje odabrati.

Neka je $\mathbf{z} \equiv$ (1, x_1, x_2, ..., x_{K-1}, z_1, z_2, ..., z_M) i L=K+M. Može se pokazati da pod pretpostavkama da je:
E(**z**'u) = 0, E(**z**'**z**) = L, E(**z**'**x**) = K i E(u^2**z**'**z**) = E(u^2)E(**z**'**z**), **2SLS (two-stage least squares)** metoda daje najefikasnijeg IV procjenitelja. Dokaz se nalazi u poglavlju 5.2.3 knjige [8]. Preciznije, od svih linearnih kombinacija elemenata u **z**, 2SLS metoda odabire onu koja je najkoreliranija sa x_K. To znači da za egzogenu varijablu x_j najbolji instrument je ta sama varijabla.

Napišimo jednadžbu u reduciranoj formi za x_K kao:

$$x_K = \delta_0 + \delta_1 x_1 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + ... + \theta_M z_M + r_K \tag{1.29}$$

gdje je po definiciji očekivanje varijable r_K jednako nula i r_K je nekorelirana sa svim varijablama na desnoj strani jednadžbe. Ako sa x_K^* označimo:

$$x_K^* = \delta_0 + \delta_1 x_1 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + ... + \theta_M z_M \tag{1.30}$$

onda je x_K^* nekorelirana sa u kao linearna kombinacija elemenata nekoreliranih sa u. Kada bismo mogli opaziti x_K^*, onda bismo nju uzeli za instrumentalnu varijablu za x_K u jednadžbi (1.20). Problem je što su koeficijenti δ_j i θ_j populacijski parametri, no uz pretpostavku da nema linearne zavisnosti između egzogenih varijabli, konzistentno ih možemo procijeniti korištenjem OLS metode, gledajući svaku opservaciju i = 1, 2, ..., N:

$$\hat{x}_{iK} = \hat{\delta}_0 + \hat{\delta}_1 x_{i1} + ... + \hat{\delta}_{K-1} x_{i,K-1} + \hat{\theta}_1 z_{i1} + ... + \hat{\theta}_M z_{i,M} \tag{1.31}$$

Za svaki i definiramo $\hat{\mathbf{x}}_i \equiv (1, x_{i1}, x_{i2}, ..., x_{i,K-1}, \hat{x}_{iK})$. Korištenjem $\hat{\mathbf{x}}_i$ kao instrumenta za $\mathbf{x}_i$, iz formule (1.27) slijedi:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \hat{\mathbf{x}}_i' \mathbf{x}_i \right)^{-1} \left(\sum_{i=1}^{N} \hat{\mathbf{x}}_i' y_i \right) \tag{1.32}$$

Dakle, **prvi korak** je regresija oblika:

$$x_K \; na \; 1, x_1, ..., x_{K-1}, z_1, ..., z_M \tag{1.33}$$

po svim opservacijama i.

Drugi korak je regresija oblika:

$$y \; na \; 1, x_1, ..., x_{K-1}, \hat{x}_K \tag{1.34}$$

kako bismo dobili $\hat{\beta}_j$.

1.4.3 Korištenje instrumentalnih varijabli za procjenu prosječnog efekta tretmana

Metodu instrumentalnih varijabli (IV) koristimo kada nismo sigurni vrijedi li nezavisnost o tretmanu. IV metode mogu biti veoma korisne pri procjenjivanju ATE, ukoliko imamo dostupnu dobru instrumentalnu varijablu. Kao i prije, korisno je y napisati kao:

$$y = \mu_0 + (\mu_1 - \mu_0)w + \upsilon_0 + w(\upsilon_1 - \upsilon_0) \tag{1.35}$$

Sada ne pretpostavljamo da je E(υ_i | w, $\mathbf{x}$) = E(υ_i | w), i=0,1. Umjesto toga, pretpostavit ćemo da se u vektoru $\mathbf{z}$ nalaze moguće instrumentalne varijable, koje su odvojene od kovarijabli, tj, $\mathbf{x}$ i $\mathbf{z}$ nemaju ništa u presjeku. U mnogim slučajevima $\mathbf{z}$ neće biti vektor, već skalar. Ako vrijedi da je $\upsilon_1 = \upsilon_0$, onda je ATE = ATE_1. Time dolazimo do sljedeće pretpostavke za IV metode:

PRETPOSTAVKA ATE.2: (a) U jednadžbi (1.35), υ_1=υ_0; (b) L(υ_0 | $\mathbf{x}$, $\mathbf{z}$) = L(υ_0 | $\mathbf{x}$); i (c) L(w | $\mathbf{x}$, $\mathbf{z}$) $\neq$ L(w | $\mathbf{x}$).

Iz (a) i (b) dijela pretpostavke ATE.2, slijedi da je:
ATE = E(y_1 - y_0) = E(μ_1 - μ_0 + υ_1 - υ_0) = E(μ_1 - μ_0) = μ_1 - μ_0, pa stoga dodavanjem

$\pm$L(υ_0 | $\mathbf{x}$, $\mathbf{z}$) = $\pm$L(υ_0 | $\mathbf{x}$) = $\pm\mathbf{x}\boldsymbol{\beta}$ u jednadžbu (1.35) dobivamo:

$$y = \delta_0 + \alpha w + \mathbf{x}\boldsymbol{\beta} + u_0 \tag{1.36}$$

gdje je α = ATE, a u_0 = υ_0 - L(υ_0 | $\mathbf{x}$, $\mathbf{z}$).

Iz svojstava linearne projekcije slijedi da je E(u_0) = 0 i u_0 je nekorelirana sa ($\mathbf{x}$, $\mathbf{z}$). Pod pretpostavkom ATE.2, α (i drugi parametri jednadžbe (1.36)) mogu biti konzistentno procijenjeni 2SLS metodom.

Pretpostavka ATE.2 (b) vrijedi ako je vektor instrumenata $\mathbf{z}$ nezavisan od (y_0, $\mathbf{x}$). Na primjer, neka je z skalar koji određuje prikladnost osobe za program obuke za posao. Sudjelovanje u programu, w, moglo bi biti korelirano s nekom neopazivom varijablom koja se nalazi u υ_0, kao što je sposobnost osobe. Ako je prikladnost na slučajan način pridružena osobama iz uzorka, možemo reći da je z nezavisna od (y_0, $\mathbf{x}$). Budući da sposobnost pozitivno utječe na sudjelovanje u programu, pretpostavka ATE.2 (c) vrijedi. Ipak, slučajno pridruživanje prikladnosti ne znači nužno da je prikladnost dobar instrument za sudjelovanje, što vidimo iz sljedećeg primjera:

Neka je z varijabla koja predstavlja broj izvučen na lotu i neka taj broj utječe na regrutiranje osobe za vojsku. Preciznije, neka je z instrument za regrutaciju. Očito je da pretpostavka ATE.2 (c) vrijedi. Osobe s manjim izvučenim brojem imaju veće šanse biti odabrane za vojsku, stoga bi osoba s manjim brojem mogla odlučiti namjerno produžiti školovanje kako bi dobila odgodu od vojske. Ako bi se u obzir (u vektor $\mathbf{x}$) uključio i ukupan broj godina školovanja (kako osobe koje se "predugo" školuju više ne bi mogle dobivati odgodu), taj problem bio bi riješen. S druge strane, osoba s većim brojem vjerojatno neće biti odabrana za vojsku, pa je moguće da njezin poslodavac odluči investirati u nju i poslati je na dodatnu obuku (doškolovanje). Toj osobi sada raste vjerojatnost za regrutaciju, jer ona povećava svoj broj godina školovanja. Iz toga vidimo da uključivanje godina školovanja u $\mathbf{x}$ nije efikasno rješenje.

Pretpostavka ATE.2 (b) ne znači da $\mathbf{z}$ ne može biti koreliran s elementima $\mathbf{x}$-a. Na primjer, neka je w binarna varijabla koja označava posjeduje li osoba fakultetsku diplomu, a z instrument koji označava udaljenost srednje škole koju je osoba pohađala do najbližeg fakulteta. Tada je moguće da je z koreliran s elementima $\mathbf{x}$, kao što su, na primjer, geografski indikatori.

Pod pretpostavkama ATE.2, 2SLS metoda primijenjena na jednadžbu (1.36) daje konzistentnu i asimptotski normalnu procjenu. No, pod jačom pretpostavkom možemo dobiti još efikasnijeg IV procjenitelja.

PRETPOSTAVKA ATE.2’: (a) U jednadžbi (1.35), $\upsilon_1 = \upsilon_0$; (b) $E(\upsilon_0 \mid \mathbf{x}, \mathbf{z}) = L(\upsilon_0 \mid \mathbf{x})$; (c) $P(w=1 \mid \mathbf{x}, \mathbf{z}) \neq P(w=1 \mid \mathbf{x})$, $P(w=1 \mid \mathbf{x}, \mathbf{z}) = G(\mathbf{x}, \mathbf{z}; \boldsymbol{\gamma})$ u parametarskom obliku, najčešće logit ili probit; te (d) $\mathrm{Var}(\upsilon_0 \mid \mathbf{x}, \mathbf{z}) = \sigma_0^2$.

U jednadžbi (1.36) za grešku u_0 vrijedilo je $u_0 = \upsilon_0 - L(\upsilon_0 \mid \mathbf{x}, \mathbf{z})$. Sada iz (a) i (b) dijela pretpostavke ATE.2’ vidimo da je $E(u_0 \mid \mathbf{x}, \mathbf{z}) = 0$. Pretpostavka ATE.2’ (d) implicira da je i $\mathrm{Var}(u_0 \mid \mathbf{x}, \mathbf{z})$ konstanta. U poglavlju 14.5.3 knjige [8] dokazano je da je optimalna IV za w upravo $E(w \mid \mathbf{x}, \mathbf{z}) = G(\mathbf{x}, \mathbf{z}; \boldsymbol{\gamma})$. Stoga možemo koristiti IV metodu u 2 koraka koju ćemo nazvati Procedura 1.

Procedura 1:

1. Metodom MLE procijenimo $P(w=1 \mid \mathbf{x}, \mathbf{z}) = G(\mathbf{x}, \mathbf{z}; \boldsymbol{\gamma})$ kako bismo dobili vjerojatnosti $\hat{G}_i$. Ako je $G(\cdot)$ funkcija distribucije standardne normalne slučajne varijable, radi se o probit modelu.

2. Procijenimo jednadžbu (1.36) koristeći instrumente 1, $\hat{G}_i$ i $\mathbf{x}_i$.

Primjer 1.4.1. *(Procjena utjecaja obrazovanja na plodnost)*

Promatramo utjecaj obrazovanja, trajanja minimalno sedam godina, na plodnost. Podaci koje imamo obuhvaćaju žene u fertilnoj dobi u Bocvani. Oko 21 posto žena nema niti jednu godinu obrazovanja. Oko 33 posto ih ima sedam godina obrazovanja. Neka varijabla y označava broj živuće djece određene osobe i neka je w binarni indikator za pohađanje obrazovanja tokom minimalno sedam godina. Elementi $\mathbf{x}$ su *dob*, *dob*^{2}, *brak* (mjeri se je li osoba ikad bila u braku), *urban* (živi li osoba u urbanom području), *električna energija* (ima li osoba električnu energiju) i *tv* (ima li osoba televiziju).

OLS procjena daje procjenu za ATE = -.394 (se = .050). Za varijablu w koristimo instrumentalnu varijablu *prvapol* koja je binarna i označava je li osoba rođena u prvoj polovici godine. Iz podataka je vidljivo da postoji značajna negativna koreliranost varijabli w i *prvapol*. IV metoda (2SLS) daje puno veću procjenu za ATE od čak -1.131, no značajnost toga nije velika budući je se = .619. Procjena iz Procedure 1 još je jačeg iznosa i velike značajnosti. ATE iznosi -1.975, a se = .332. Možemo zaključiti da osoba koja je obrazovanija ima manji broj djece. Problem je što se vrijednosti y grupiraju oko nule (većina promatranih žena nema djece). Stoga je moguće da pretpostavke (a) i (b) iz ATE.2' ne vrijede, pa i da je Procedura 1 precijenila efekt obrazovanja na broj djece.

Ako vrijede prve tri pretpostavke u ATE.2', tada je (zbog G je instrument za w i E(u_0 | $\mathbf{x}$, $\mathbf{z}$) = 0):

$$E(\,y \mid \mathbf{x}, \mathbf{z}) = \delta_0 + \alpha G(\mathbf{x}, \mathbf{z}; \boldsymbol{\gamma}) + \mathbf{x}\boldsymbol{\beta} \tag{1.37}$$

Korištenje prilagođenih vjerojatnosti, dobivenih u prvom koraku iz probit ili logit modela, kao instrumenta za w dobar je način za iskoristiti "binarnu prirodu" endogene eksplanatorne varijable. No, ključna pretpostavka bila je da E(υ_0 | $\mathbf{x}$, $\mathbf{z}$) ovisi samo o $\mathbf{x}$ i da je linearno u $\mathbf{x}$, što je jača pretpostavka od pretpostavke ATE2. (b). Također, pretpostavljali smo da je υ_1=υ_0.

Ako je $\upsilon_1 \neq \upsilon_0$, postoji interakcija $w(\upsilon_1$-$\upsilon_0)$ u jednadžbi (1.35). Tada, u većini slučajeva, IV procjenitelj, koji koristi instrumente $\mathbf{z}$ ili $\hat{G}$ za w, neće konzistentno procijeniti ATE. Stoga je potrebno postaviti neke dodatne uvjete. Neka je:

$$E(\upsilon_0 \mid \mathbf{x}, \mathbf{z}) = E(\upsilon_0 \mid \mathbf{x}) \quad i \quad E(\upsilon_1 \mid \mathbf{x}, \mathbf{z}) = E(\upsilon_1 \mid \mathbf{x}). \tag{1.38}$$

Jednadžbu (1.35) tada možemo napisati kao:

$$y = \mu_0 + \alpha w + g_0(\mathbf{x}) + w[g_1(\mathbf{x}) - g_0(\mathbf{x})] + e_0 + w(e_1 - e_0) \tag{1.39}$$

gdje je α=ATE, $\upsilon_i = g_i(\mathbf{x}) + e_i$ i E(e_i | $\mathbf{x}$, $\mathbf{z}$) = 0, za i= 0, 1.

Prethodnu jednadžbu možemo procijeniti IV metodom, gdje je greška modela $e_0 + w(e_1 - e_0)$. Funkcije g_0 i g_1 tipično su linearne u parametrima. Konkretnije, neka je

$$g_0(\mathbf{x}) = \eta_0 + \mathbf{x}\boldsymbol{\beta}_0, \quad g_1(\mathbf{x}) - g_0(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta} \tag{1.40}$$

gdje je $\boldsymbol{\delta}$ = E($\mathbf{x}$). Ako takve g_1 i g_0 uvrstimo u jednadžbu (1.39) dobivamo:

$$y = \gamma + \alpha w + \mathbf{x}\boldsymbol{\beta}_0 + w(\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta} + e_0 + w(e_1 - e_0) \tag{1.41}$$

Vidimo da je potrebno pronaći instrumente za w i $w(\mathbf{x}\text{-} \boldsymbol{\psi})$ kako bismo imali varijable nekorelirane s greškom modela. Ukoliko je $q \equiv q(\mathbf{x},\mathbf{z})$ instrument za w, prirodno je za instrument za $w \cdot \mathbf{x}$ uzeti $q \cdot \mathbf{x}$. Tada, ako je q efikasan za w, i $q \cdot \mathbf{x}$ će biti efikasan za $w \cdot \mathbf{x}$. Ako vrijedi da je

$$e_1 = e_0, \tag{1.42}$$

greška modela je samo e_0 i zbog uvjeta $\mathrm{E}(e_0 \mid \mathbf{x}, \mathbf{z}) = 0$ IV procjena jednadžbe (1.41) daje konzistentnu i asimptotski normalnu procjenu.

PRETPOSTAVKA ATE.3: Neka je y definiran kao u jednadžbi (1.35) i neka vrijede pretpostavke (1.38), (1.40), (1.42) i ATE.2'.

Tada možemo definirati Proceduru 2:

Procedura 2:
Procijenimo jednadžbu:

$$y = \gamma + \alpha w_i + \mathbf{x}_i \boldsymbol{\beta}_0 + w_i(\mathbf{x}_i - \bar{\mathbf{x}})\boldsymbol{\delta} + error_i \tag{1.43}$$

korištenjem instrumenata 1, $\hat{G}_i$, $\mathbf{x}_i$ i $\hat{G}_i(\mathbf{x}_i - \bar{\mathbf{x}})$, tj. kao i ranije, $\hat{G}_i$ je instrument za w_i.

Ako dodatno vrijedi pretpostavka ATE.2' (d), gornja procedura daje efikasnog procjenitelja. Alternativno, umjesto procjene G_i, možemo uzeti $\mathbf{z}_i$ i interakcije $\mathbf{z}_i$ i $\mathbf{x}_i$ kao instrumente.

Primjer 1.4.2. *(Evaluacija obuke za posao pomoću IV metode)*

Pretpostavimo da vektor $\mathbf{x}$ sadržava podatke o obrazovanju, iskustvu i iskustvu na kvadrat. Neka z označava podobnost za sudjelovanje u programu i neka je w indikator sudjelovanja u programu. Tada možemo procijeniti sljedeću jednadžbu:

$$\log(nadnica) = \mu_0 + \alpha w + \beta_{01}\ obrazovanje + \beta_{02}\ iskustvo + \beta_{03}\ iskustvo^2$$

$$+ \delta_1 w\ (obrazovanje\text{-}\ \overline{obrazovanje}) + \delta_2 w\ (iskustvo\text{-}\ \overline{iskustvo}) + \delta_3 w\ (iskustvo^2\text{-}\ \overline{iskustvo^2})$$

$$+\ error$$

koristeći IV metodu i instrumente: 1, z, $obrazovanje$, $iskustvo$, $iskustvo^2$ te interakcije z s kovarijablama kojima smo oduzeli očekivanje.

Za provedbu Procedure 2 važne pretpostavke bile su (1.40) i (1.42). Ipak, uvjet (1.42) može se modificirati sa:

$$E[w(e_1 - e_0) \mid \mathbf{x}, \mathbf{z}] = E[w(e_1 - e_0)] \tag{1.44}$$

Ako vrijede (1.38), (1.40) i (1.44), IV procjenitelj je konzistentan. Naime, dodajući $\pm$E[$w(e_1 - e_0)$] u jednadžbu (1.41) dobijemo:

$$y = \xi + \alpha w + \mathbf{x}\boldsymbol{\beta}_0 + w(\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta} + e_0 + r \tag{1.45}$$

gdje je $r = w(e_1 - e_0)$ - E[$w(e_1 - e_0)$], $\xi = \gamma$ + E[$w(e_1 - e_0)$]. Pod uvjetom (1.44) je E(r | **x**, **z**) = 0, pa je zbog uvjeta u (1.39) i E(e_0 - r | **x**, **z**) = 0. Stoga, svaka funkcija od (**x**, **z**) može biti korištena kao instrument u jednadžbi (1.45). Ako modificiramo pretpostavku ATE.3 na sljedeći način, procedura 2 i dalje ostaje konzistentna:

PRETPOSTAVKA ATE.3’: Neka je y definiran kao u jednadžbi (1.35) i neka vrijede pretpostavke (1.38), (1.40) i (1.44). Također neka pretpostavka ATE.2’ (c) vrijedi.

Ako je **z** nezavisan od $(y_0, y_1, \mathbf{x})$, kovarijable mogu u potpunosti biti izbačene iz analize, vodeći do IV procjene jednostavne regresijske jednadžbe $y = \xi + \alpha w + error$, što je pokazao Angrist (1991) u članku [1].

Pretpostavimo da je

$$E[w \mid \mathbf{x}, \mathbf{z}, e_1 - e_0] = h(\mathbf{x}, \mathbf{z}) + k(e_1 - e_0) \tag{1.46}$$

za neke funkcije $h(\cdot)$ i $k(\cdot)$ i da je

$$e_1 - e_0 \quad nezavisan\ od \quad (\mathbf{x}, \mathbf{z}). \tag{1.47}$$

Tada vrijedi:

$$\begin{aligned} E[w(e_1 - e_0) \mid \mathbf{x}, \mathbf{z}] &= (LIE) = E[\ E[w(e_1 - e_0) \mid \mathbf{x}, \mathbf{z}, e_1 - e_0] \mid \mathbf{x}, \mathbf{z}] \\ &= h(\mathbf{x}, \mathbf{z})\ E(e_1 - e_0 \mid \mathbf{x}, \mathbf{z})\ +\ E[\ (e_1 - e_0)k(e_1 - e_0) \mid \mathbf{x}, \mathbf{z}] \\ &= h(\mathbf{x}, \mathbf{z}) \cdot 0\ +\ E[\ (e_1 - e_0)k(e_1 - e_0)\ ] \\ &= E[\ (e_1 - e_0)k(e_1 - e_0)\ ] \end{aligned} \tag{1.48}$$

budući da je $E(e_1 - e_0 \mid \mathbf{x}, \mathbf{z}) = 0$.

Zbog nezavisnosti $e_1 - e_0$ i ($\mathbf{x}$, $\mathbf{z}$), vrijedi i $E(e_1 - e_0) = 0$. Dakle, ako bi funkcija k bila identiteta, gornja jednadžba predstavljala bi $\mathrm{Var}(e_1 - e_0)$. U pravilu, pretpostavka o nezavisnosti $(e_1 - e_0)$ i ($\mathbf{x}$, $\mathbf{z}$) vrijedi ako je y neprekidna, no ne i ako je diskretna varijabla.

Pretpostavka (1.46) narušena je, na primjer, kod probit modela

$$P(w = 1 \mid \mathbf{x}, \mathbf{z}, e_1 - e_0) = \Phi[\pi_0 + \mathbf{x}\boldsymbol{\pi}_1 + \mathbf{z}\boldsymbol{\pi}_2 + \rho(e_1 - e_0)] \tag{1.49}$$

kod kojeg nije moguće odijeliti ($\mathbf{x}$, $\mathbf{z}$) i $(e_1 - e_0)$. No čak i ako sama pretpostavka ne vrijedi, može predstavljati dobru aproksimaciju.

Sada ćemo pretpostaviti da vrijede (1.47), (1.49) i dodatna pretpostavka da je

$$e_1 - e_0 \sim N(0, \tau^2). \tag{1.50}$$

Iz te tri pretpostavke slijedi da je

$$P(w = 1 \mid \mathbf{x}, \mathbf{z}, e_1 - e_0) = \Phi[\theta_0 + \mathbf{x}\boldsymbol{\theta}_1 + \mathbf{z}\boldsymbol{\theta}_2] \tag{1.51}$$

gdje je svaki *theta* jednak odgovarajućem *pi* pomnoženom sa $(1 + \rho^2\tau^2)^{-1/2}$. Neka a označava skrivenu grešku prethodne jednadžbe koja je standardna normalna, te definiramo $c \equiv e_1 - e_0$. Tada iz uvjeta (1.47), (1.49) i (1.50) slijedi da je vektor (a, c) iz bivarijantne distribucije s očekivanjem nula i nezavisan je od ($\mathbf{x}$, $\mathbf{z}$). Stoga vrijedi:

$$E(\, c \mid a, \mathbf{x}, \mathbf{z}) = (zbog\ (1.47)) = E(\, c \mid a) = \xi a \tag{1.52}$$

i

$$E(\, w, c \mid \mathbf{x}, \mathbf{z}) = E[\, E(\, w, c \mid a, \mathbf{x}, \mathbf{z}) \mid \mathbf{x}, \mathbf{z}\,] = \xi E(\, wa \mid \mathbf{x}, \mathbf{z}) \tag{1.53}$$

Zbog toga što je a ∼ N(0,1) i nezavisan od ($\mathbf{x}$, $\mathbf{z}$) je:

$$E(\, w, a \mid \mathbf{x}, \mathbf{z}) = \int_{-\infty}^{\infty} \mathbf{1}[\theta_0 + \mathbf{x}\boldsymbol{\theta}_1 + \mathbf{z}\boldsymbol{\theta}_2 + a \geq 0]\, a\, \phi(a)\, da = \phi(-(\theta_0 + \mathbf{x}\boldsymbol{\theta}_1 + \mathbf{z}\boldsymbol{\theta}_2)) = \phi(\theta_0 + \mathbf{x}\boldsymbol{\theta}_1 + \mathbf{z}\boldsymbol{\theta}_2) \tag{1.54}$$

gdje je $\phi(\cdot)$ standardna normalna funkcija gustoće. Dakle, pokazali smo da je:

$$E(\, w, c \mid \mathbf{x}, \mathbf{z}) = \xi\, \phi(\theta_0 + \mathbf{x}\boldsymbol{\theta}_1 + \mathbf{z}\boldsymbol{\theta}_2) \tag{1.55}$$

Uvrštavanjem u (1.41) slijedi:

$$y = \gamma + \alpha w + \mathbf{x}\boldsymbol{\beta}_0 + w(\mathbf{x} - \boldsymbol{\psi})\boldsymbol{\delta} + e_0 + \xi\, \phi(\theta_0 + \mathbf{x}\boldsymbol{\theta}_1 + \mathbf{z}\boldsymbol{\theta}_2) + r \tag{1.56}$$

gdje je $r = wc$ - $E(\, w, c \mid \mathbf{x}, \mathbf{z})$. Greška ovog modela ima očekivanje nula uvjetno na ($\mathbf{x}$, $\mathbf{z}$) pa se parametri modela procijene koristeći IV metodu.

PRETPOSTAVKA ATE.4: Neka je y definiran kao u jednadžbi (1.35) i neka vrijede pretpostavke (1.38), (1.40), (1.47), (1.49) (sa $\boldsymbol{\pi}_2 \neq 0$) i (1.50).

Procedura 3:

1. Procijenimo θ_0, $\boldsymbol{\theta}_1$ i $\boldsymbol{\theta}_2$ iz probit modela w-a na (1, $\mathbf{x}$, $\mathbf{z}$) i formiramo dobivene vjerojatnosti $\hat{\Phi}_i$ zajedno sa $\hat{\phi}_i = \phi(\theta_0 + \mathbf{x}_i\boldsymbol{\theta}_1 + \mathbf{z}_i\boldsymbol{\theta}_2\,)$ za $i = 1, 2, \ldots$, N.

2. Procijenimo jednadžbu:

$$y = \gamma + \alpha w_i + \mathbf{x}_i\, \boldsymbol{\beta}_0 + w_i(\mathbf{x}_i - \bar{\mathbf{x}})\boldsymbol{\delta} + \xi\hat{\phi}_i + error_i \tag{1.57}$$

korištenjem instrumenata 1, $\hat{\Phi}_i$, $\mathbf{x}_i$, $\hat{\Phi}_i(\mathbf{x}_i$ - $\bar{\mathbf{x}})$, $\hat{\phi}_i$.

Čak i kada je $\xi \neq 0$, efekt dodavanja $\hat{\phi}_i$ procijenjenom α je jako mali. Razmotrit ćemo jednadžbu (1.56) bez kovarijabli $\mathbf{x}$ i sa instrumentom z koji je skalar:

$$y = \gamma + \alpha w + \xi\ \phi(\theta_0 + \theta_1 z) + u, \qquad E(u \mid z) = 0 \tag{1.58}$$

IV procjenitelj za α dobijemo izostavljanjem $\phi(\theta_0 + \theta_1 z)$. Ako je z instrument za w, pod pretpostavkom da su z i $\phi(\theta_0 + \theta_1 z)$ nekorelirani, taj procjenitelj je konzistentan. Naime, iako je $\phi(\theta_0 + \theta_1 z)$ funkcija od z, te dvije varijable mogu imati malu koreliranost budući je z monotona, a $\phi(\theta_0 + \theta_1 z)$ simetrična oko $-(\theta_0/\theta_1)$. Problem nastaje jedino ako je z binarna varijabla. Tada $\phi(\theta_0+\theta_1 z)$ poprima samo dvije vrijednosti. To znači da je u linearnoj relaciji sa z.

Za y definiran kao u (1.35) vrijedi: $ATE_1(\mathbf{x}) = \mathrm{E}(\, y_1 - y_0 \mid \mathbf{x}, w = 1) = \mu_1 - \mu_0 +$ $\mathrm{E}(\, \upsilon_1 - \upsilon_0 \mid \mathbf{x}, w = 1)$. Ako modificiramo pretpostavku ATE.3', pomoću IV metode možemo konzistentno procijeniti ATE_1.

PRETPOSTAVKA ATE.3": (a) Neka je y definiran kao u jednadžbi (1.35) i neka vrijedi da je $\mathrm{E}(\, \upsilon_0 \mid \mathbf{x}, \mathbf{z}) = \mathrm{E}(\, \upsilon_0 \mid \mathbf{x})$; (b) $\mathrm{E}(\, \upsilon_1 - \upsilon_0 \mid \mathbf{x}, \mathbf{z}, w = 1) = \mathrm{E}(\, \upsilon_1 - \upsilon_0 \mid \mathbf{x}, w = 1)$; i (c) pretpostavka ATE.2' (c) vrijedi.

Pretpostavka ATE.3" (a) narušena je ako sudionici mijenjaju svoje ponašanje ovisno o $\mathbf{z}$. Sjetimo se da je $\upsilon_1 - \upsilon_0$ korist od sudjelovanja, specifična za svaku osobu. Pretpostavka

ATE.3" (b) znači da za grupu koja je podvrgnuta tretmanu i za koju je fiksiran $\mathbf{x}$, korist nije predvidiva uvjetno na $\mathbf{z}$. U primjeru s loto brojevima i regrutacijom varijabla z označavala je broj izvučen na lotu. Osoba s većim z imala je veću šansu izbjeći vojsku. Ipak, i osobe s "velikim" z mogle su se dobrovoljno prijaviti za vojsku. Mogli bismo reći da je za te osobe odlazak u vojsku imao "veći značaj" nego što je to bio slučaj kod osoba koje su se također dobrovoljno prijavile za vojsku, no imale su manji z. Dakle, za osobe koje se dobrovoljno žele prijaviti za vojsku, $v_1 - v_0$ i z su pozitivno korelirani. Pretpostavka ATE.3" (b) je moguća u situacijama kada je z binarna varijabla koja označava podobnost za sudjelovanje u programu i kada je ta podobnost slučajno raspoređena po jedinkama te utječe samo na odluku o sudjelovanju, ne i na druge promjene u ponašanju.

Dodavanjem ± E($v_1 - v_0 \mid \mathbf{x}, w = 1$) u (1.35) dobivamo:

$$\begin{aligned} y =& \mu_0 + g_0(\mathbf{x}) + w[\ (\mu_1 - \mu_0) - E(v_1 - v_0 \mid \mathbf{x}, w = 1)\ ] \\ &+ w[\ (v_1 - v_0) - +E(v_1 - v_0 \mid \mathbf{x}, w = 1)\ ] + e_0 \\ &= \mu_0 + g_0(\mathbf{x}) + w \cdot ATE_1 + a + e_0 \end{aligned} \tag{1.59}$$

gdje je $a \equiv w[\ (v_1 - v_0) - E(v_1 - v_0 \mid \mathbf{x}, w = 1)\ ]$, $v_0 = g_0(\mathbf{x}) + e_0$ i $E(e_0 \mid \mathbf{x}, \mathbf{z}) = 0$.

Kada je $w = 0$, onda je i $a = 0$. Stoga, da bismo pokazali da je $E(a \mid \mathbf{x}, \mathbf{z}) = 0$, dovoljno je pokazati da je $E(a \mid \mathbf{x}, \mathbf{z}, w = 1) = 0$, budući da je $E(a \mid \mathbf{x}, \mathbf{z}) = P(w = 0) \cdot E(a \mid \mathbf{x}, \mathbf{z}, w = 0) + P(w = 1) \cdot E(a \mid \mathbf{x}, \mathbf{z}, w = 1)$. Računamo:

$$E(a \mid \mathbf{x}, \mathbf{z}, w = 1) = E(v_1 - v_0 \mid \mathbf{x}, \mathbf{z}, w = 1) - E[E(v_1 - v_0 \mid \mathbf{x}, w = 1) \mid \mathbf{x}, \mathbf{z}, w = 1]$$

$$= (\ ATE.3''b\ ) = E(v_1 - v_0 \mid \mathbf{x}, \mathbf{z}, w = 1) - E(v_1 - v_0 \mid \mathbf{x}, w = 1) = 0$$

Ako definiramo $r \equiv a + e_0$, $g_0(\mathbf{x}) = \eta_0 + \mathbf{h}(\mathbf{x})\boldsymbol{\beta}_0$ i $ATE_1(\mathbf{x}) = \tau + \mathbf{f}(\mathbf{x})\boldsymbol{\delta}$, za $\mathbf{h}(\mathbf{x})$ i $\mathbf{f}(\mathbf{x})$ vektorske funkcije, y možemo napisati kao:

$$y = \gamma_0 + \mathbf{h}_0(\mathbf{x})\boldsymbol{\beta}_0 + \tau w + w[\mathbf{f}(\mathbf{x})\boldsymbol{\delta}] + r \tag{1.60}$$

i vrijedi: $E(\ r \mid \mathbf{x}, \mathbf{z}) = E(\ a \mid \mathbf{x}, \mathbf{z}) + E(\ e_0 \mid \mathbf{x}, \mathbf{z}) = 0$.

Svi parametri jednadžbe (1.60) mogu se konzistentno procijeniti IV metodom, koristeći bilo koje funkcije od ($\mathbf{x}$, $\mathbf{z}$) kao instrumente (što uključuje i funkcije 1, $\mathbf{h}_0(\mathbf{x})$, $G(\mathbf{x}, \mathbf{z}; \hat{\boldsymbol{\gamma}})$ i $G(\mathbf{x}, \mathbf{z}; \hat{\boldsymbol{\gamma}}) \cdot \mathbf{f}(\mathbf{x})$). Prosječan efekt tretmana za tretiranu jedinku, za svaki $\mathbf{x}$, procijenjen je sa $\hat{\tau} + \mathbf{f}(\mathbf{x})\hat{\delta}$. Gledajući prosjek za sve jedinke sa $w_i = 1$ dobivamo konzistentnog procjenitelja za ATE_1.

1.4.4 Procjena lokalnog prosječnog efekta tretmana (LATE-a)

Kao i prije, neka je w binarna varijabla koja označava tretman i neka je y_1 ishod s tretmanom, a y_0 bez tretmana. Ishod y tada možemo napisati (kao i u (1.3)):

$$y = (1 - w)y_0 + wy_1 = y_0 + w(y_1 - y_0) \tag{1.61}$$

Neka je z instrumentalna varijabla. Ovdje promatramo najjednostavniji slučaj, kada je z binarna. Svaka jedinka i imat će "svoj" z_i, jednak nula ili jedan. Ovisno o tome je li z = 0 ili z = 1, toj jedinki pridružit ćemo w_0, odnosno w_1 redom. Dakle, za svaku jedinku moguće je opaziti samo jedan od w_i, i = 0,1. Status tretmana w sada možemo zapisati kao:

$$w = (1 - z)w_0 + zw_1 = w_0 + z(w_1 - w_0) \tag{1.62}$$

Na primjer, z može označavati prikladnost osobe za sudjelovanje u programu. No, w označava stvarno sudjelovanje u programu i vrijednosti te dvije varijable općenito se ne podudaraju.

Kad uvrstimo (1.62) u (1.61), dobijemo:

$$y = y_0 + w_0(y_1 - y_0) + z(w_1 - w_0)(y_1 - y_0) \tag{1.63}$$

Važna pretpostavka je da je z nezavisan od (y_0, y_1, w_0, w_1), budući da tada sva očekivanja od (y_0, y_1, w_0, w_1) uvjetno na z zapravo ne ovise o z. Dakle:

$$E(\, y \mid z = 1) = E(y_0) + E[w_0(y_1 - y_0)] + E[(w_1 - w_0)(y_1 - y_0)] \tag{1.64}$$

$$E(\, y \mid z = 0) = E(y_0) + E[w_0(y_1 - y_0)] \tag{1.65}$$

Oduzimanjem posljednje dvije jednadžbe dobivamo:

$$E(\, y \mid z = 1) - E(\, y \mid z = 0) = E[(w_1 - w_0)(y_1 - y_0)] \tag{1.66}$$

Zbog toga što je

$$E(y) = p_1 \cdot E(\, y \mid \mathbf{x} = \mathbf{c}_1) + ... + p_m \cdot E(\, y \mid \mathbf{x} = \mathbf{c}_m)$$

za $p_i = P(\mathbf{x} = \mathbf{c}_i)$, (1.66) može se zapisati kao:

$$\begin{aligned}
E[(w_1 - w_0)(y_1 - y_0)] &= 1 \cdot E(\, y_1 - y_0 \mid w_1 - w_0 = 1) \cdot P(w_1 - w_0 = 1) \\
&+ (-1) \cdot E(\, y_1 - y_0 \mid w_1 - w_0 = -1) \cdot P(w_1 - w_0 = -1) \\
&+ 0 \cdot E(\, y_1 - y_0 \mid w_1 - w_0 = 0) \cdot P(w_1 - w_0 = 0) \\
&= E(\, y_1 - y_0 \mid w_1 - w_0 = 1) \cdot P(w_1 - w_0 = 1) \\
&- E(\, y_1 - y_0 \mid w_1 - w_0 = -1) \cdot P(w_1 - w_0 = -1) \qquad (1.67)
\end{aligned}$$

Sada ćemo pretpostaviti da vrijedi:

$$w_1 \geq w_0 \qquad (1.68)$$

To svojstvo Imbens i Angrist (vidi [2]) nazvali su **monotonost** (*eng. monotonicity*) i time smo eliminirali slučaj kada je $w_1 = 0$ i $w_0 = 1$. Naime, kada je $w_1 = 0$ i $w_0 = 1$, jednadžba (1.62) postaje:

$$w = (1 - z) \cdot 1 + 0 = 1 - z \qquad (1.69)$$

tj. vrijedi ($w = 0 \Leftrightarrow z = 1$) pa time i ($w = 1 \Leftrightarrow z = 0$). Uz pretpostavku (1.68) dobivamo da će osobe koje koje nisu prikladne za sudjelovanje u programu (tj. za koje je $z = 0$), a ipak sudjeluju (tj. za njih je $w = 1$), sudjelovati ($w = 1$) i ako bi vrijednost njihovog z bila 1. Matematički, za jedinku i vrijedit će da ako ($z_i = 0 \Rightarrow w_i = 1$), onda ($z_i = 1 \Rightarrow w_i = 1$). Jedinke koje ne zadovoljavaju uvjet (1.68) nazivamo *defiers* - oni koji prkose.

Ako vrijedi pretpostavka (1.68), P($w_1 - w_0$ = -1) = 0 pa uvrštavanjem u (1.66) i (1.67) dobivamo:

$$E(\, y \mid z = 1) - E(\, y \mid z = 0) = E(\, y_1 - y_0 \mid w_1 - w_0 = 1) \cdot P(w_1 - w_0 = 1) \qquad (1.70)$$

Sada možemo definirati **LATE** kao:

$$\textbf{LATE} = E(\, y_1 - y_0 \mid w_1 - w_0 = 1) \qquad (1.71)$$

U trenutnoj situaciji, $w_1 - w_0 = 1$ znači da je $w_1 = 1$, a $w_0 = 0$. Iz jednadžbe (1.62) vidimo da je $w = 0 + z \cdot 1 = z$, stoga $LATE$ daje prosječan efekt tretmana za jedinke kod kojih varijabla z određuje hoće li te jedinke sudjelovati u tretmanu. $LATE$ dakle ovisi o instrumentu z (za razliku od ATE i ATE_1 koji ovise samo o populaciji), te se promjenom instrumenta i $LATE$ mijenja. Nažalost, subpopulaciju sa $w_1 - w_0 = 1$ teško je identificirati budući da ne možemo opaziti i w_1 i w_0.

Primjer 1.4.3. *(Utjecaj pohađanja katoličke srednje škole na broj bodova na testu)*

Neka y označava broj bodova na testu, w pohađa li osoba katoličku srednju školu, a z je li učenik katolik. U ovom slučaju, tretman je dakle "pohađanje katoličke srednje škole". $LATE$ možemo izračunati iz aritmetičke sredine efekata tretmana na broj bodova na testu onih učenika koji su katolici i koji pohađaju katoličku srednju školu. Preciznije, za one pojedince kod kojih je $w = z = 1$.

Iz slučajnog uzorka lako procijenimo $E(y \mid z = 1)$ i $E(\, y \mid z = 0)$ pa ako je i $P(w_1 - w_0 = 1)$ procjenjiva i različita od nule, LATE možemo identificirati. Iz pretpostavke (1.68) slijedi da je $w_1 - w_0$ binarna varijabla, budući je $P(w_1 - w_0 = -1) = 0$. Koristeći to, dobivamo:

$$\begin{aligned} 1 \cdot P(w_1 - w_0 = 1) &= E(w_1 - w_0) = E(w_1) - E(w_0) = E(\, w \mid z = 1) - E(\, w \mid z = 0) \\ &= 1 \cdot P(\, w = 1 \mid z = 1) - 1 \cdot P(\, w = 1 \mid z = 0) \end{aligned} \tag{1.72}$$

gdje smo u prvoj i posljednjoj jednakosti koristili činjenicu da je $E(y) = p_1 \cdot E(y \mid \mathbf{x} = \mathbf{c}_1) + ... + p_m \cdot E(\, y \mid \mathbf{x} = \mathbf{c}_m)$, a u pretposljednjoj jednadžbu (1.62). Iz slučajnog uzorka na (w, z) možemo procijeniti obje uvjetne vjerojatnosti. Kako bi vrijedilo da je $P(w_1 - w_0 = 1)$ različita od nule, nužno je da je:

$$P(\, w = 1 \mid z = 1) \neq P(\, w = 1 \mid z = 0) \tag{1.73}$$

Da zaključimo, pod pretpostavkama: z je nezavisan od (y_0, y_1, w_0, w_1), (1.68) i (1.73)

$$LATE = [\, E(\, y \mid z = 1) - E(\, y \mid z = 0)\,] \;/\; [\, P(\, w = 1 \mid z = 1) - P(\, w = 1 \mid z = 0)\,] \tag{1.74}$$

Stoga je konzistentan procjenitelj za $LATE$ jednak $\widehat{LATE} = (\bar{y}_1 - \bar{y}_0)/(\bar{w}_1 - \bar{w}_0)$, gdje je $\bar{y}_1$ srednja vrijednost svih y_i po svim jedinkama i za koje je $z_i = 1$, a $\bar{y}_0$ srednja vrijednost svih y_i za koje je $z_i = 0$. Analogno definiramo i $\bar{w}_1$ i $\bar{w}_0$.

Poglavlje 2

Utjecaj IWT subvencija na R&D u Flandriji

2.1 Value for money? New microeconometric evidence on public R&D grants in Flanders

U ovom poglavlju komentirat ćemo rezultate istraživanja koje su proveli Dirk Czarnitzki i Cindy Lopes-Bento i objavili u članku naziva **Value for money? New microeconometric evidence on public R&D grants in Flanders** 2012. godine. Koristeći rezultate prijašnjih istraživanja u području inovacija i podatke prikupljene od IWT-a (Agencije za inovacije "Innovatie door Wetenschap en Technologie in Vlaanderen") testirali su utjecaj IWT subvencija na R&D u Flandriji, no za razliku od prijašnjih sličnih istraživanja, promatrali su i mnogo više od toga, kao što je stabilnost efekta tretmana tokom vremena, učinak uzastopnog ponavljanja subvencija na efekt tretmana, učinak više subvencioniranih projekata istovremeno i slično.

2.2 IWT subvencije

Agencija za inovacije "Innovatie door Wetenschap en Technologie in Vlaanderen" (IWT) osnovana je 1991. godine od strane flamanske vlade s ciljem poticanja R&D-a (*eng. research and development*) i inovacija u Flandriji - jednoj od triju pokrajina u Belgiji. Glavni cilj R&D-a je razvijanje novih proizvoda, kreiranje novih znanja i unaprjeđenje tehnologije.

IWT subvencije namijenjene su manjim i velikim kompanijama, sveučilištima i drugim inovativnim organizacijama. U prosjeku 600 kompanija godišnje dobije neku vrstu IWT subvencije. Za mala i srednja poduzeća (SME, *eng. small and medium enterprises*) kreirani su posebni KMO programi.

Ekonomski argument za vladino "uplitanje" u R&D aktivnosti privatnog sektora dao je Arrow 1962. godine (vidi [3]). Neki od njegovih zaključaka su sljedeći:

(i) Zbog toga što se svaka jedinica informacije može koristiti beskonačan broj puta, njezin trošak trebao bi iznositi nula. Iz tog razloga pojedincu se teško odlučiti da investira u nova znanja.

(ii) Dio kreiranog znanja uvijek se prelije i na druge agente, tj. niti jedna kompanija ne može u potpunosti iskoristiti koristi od R&D investicije.

(iii) Zbog nesigurnosti po pitanju uspješnosti projekta, kompanije se teško odlučuju na investicije i inovacije, pogotovo ako se radi o onima veće rizičnosti. Također, investicije u R&D, za razliku od onih u fizički kapital, ne mogu se koristiti kao kolateral pri podizanju kredita.

Zbog postojanja prelijevanja informacija i činjenice da mnogo kompanija nema dovoljno ambicije i sredstava za ulaganje u inovacije, možemo zaključiti da je subvencioniranje od strane vlade poželjno, pogotovo ako se radi o projektu velike ekonomske važnosti za društvo u cjelini. Ipak, postavlja se pitanje dovodi li javno subvencioniranje do istiskivanja privatnih investicija. Mnoga istraživanja provedena su na tu temu. Zaključci tih istraživanja su da potpunog efekta istiskivanja nema, no neka od njih dokazuju postojanje parcijalnog efekta istiskivanja.

2.3 Ekonometrijske metode

Razne metode procjene uključuju difference-in-difference procjenu, metodu instrumentalnih varijabli, neparametarsko uparivanje bazirano na vjerojatnosti sklonosti i sl. Difference-in-difference metodu ne možemo primijeniti budući da ona zahtjeva postojanje podataka prije i nakon tretmana (za svaku jedinku), što u ovom slučaju nemamo. Kao što smo već prije u radu objasnili, za korištenje metode instrumentalnih varijabli potrebno je pronaći dobar instrument, što može biti dosta zahtjevno. Prednost uparivanja je u tome što ne zahtjeva pretpostavke o funkcionalnim formama i distribuciji grešaka. Mana je što uparivanje pretpostavlja da smo opazili i uključili u model sve varijable koje određuju primanje tretmana, u ovom slučaju primanje subvencije.

Prosječan efekt tretmana tretiranih jedinki definirali smo kao:

$$ATE_1 = E(y_1 - y_0 \mid w = 1) \tag{2.1}$$

gdje w označava tretman, y_1 ishod sa, a y_0 ishod bez tretmana. U prethodnom poglavlju objasnili smo da je $E(y_1 \mid w = 1)$ moguće procijeniti, no za $E(y_0 \mid w = 1)$ to nije moguće. Budući da dodjela subvencije nije na slučajan način raspoređena, onda je $E(y_0 \mid w = 1) \neq E(y_0 \mid w = 0)$. Uvođenjem pretpostavke ATE.1', tj. pretpostavke nezavisnosti o tretmanu u smislu uvjetnog očekivanja (*"ignorability in a conditional mean independence sense"* = *CIA*), možemo riješiti problem selekcije. U našem slučaju, ta pretpostavka znači da su sudjelovanje u programu i potencijalan ishod statistički nezavisni za kompanije s istim karakteristikama, popisanima u vektoru **x**. Iz pretpostavke ATE.1' također se vidi da je:

$$E(y_0 \mid w = 1, \mathbf{x}) = E(y_0 \mid w = 0, \mathbf{x}) \tag{2.2}$$

pa ATE_1 možemo napisati kao:

$$ATE_1 = E(y_1 \mid w = 1, \mathbf{x} = \mathbf{x}_K) - E(y_0 \mid w = 0, \mathbf{x} = \mathbf{x}_K) \tag{2.3}$$

2.4 Metoda uparivanja

Metoda uparivanja koja se koristi bazirana je na vjerojatnosti sklonosti. Kompanija koja je primila subvenciju uparuje se s "najbližom" kompanijom koja nije primila subvenciju. "Bliske" kompanije imaju sličnu vrijednost vjerojatnosti sklonosti $\hat{p}(\mathbf{x})$ procijenjenu iz probit modela na binarnu varijablu w. Dodatno, zahtijeva se da su za uparene kompanije podaci iz iste godine te da je broj patenata u obje kompanije otprilike podjednak.

Također, presjek između kontrolne i tretirane grupe mora biti dovoljno velik, što se postiže na način da se u kontrolnoj grupi nađe najmanja (*min*) i najveća (*max*) vjerojatnost sklonosti, a zatim u tretiranoj grupi maknu sve opservacije kod kojih je vjerojatnost sklonosti manja od *min* i veća od *max* vrijednosti.

Kako bi se izbjegla "loša uparivanja", postavlja se prag koji označava maksimalnu dopuštenu udaljenost između uparenih jedinki. Tu ideju prvi su predstavili Cochran i Rubin (1973) (vidi [4]) i nazvali je **caliper matching**- "*metoda šestara*". Ideja je da se oko jedinke koja se uparuje "šestarom" opiše kružnica radijusa ϵ. Ta jedinka zatim se uparuje ili sa najbližom jedinkom koja je upala u krug opisan šestarom ili sa svim jedinkama koje su upale u krug. U potonjem slučaju, uzimanjem aritmetičke sredine po svim parovima, dobije se jedan par koji se koristi u daljnjoj analizi.

Sam postupak traženja para izgleda ovako: Prvo se odabere jedna opservacija iz skupa tretiranih kompanija te se izbriše iz tog skupa. Zatim se računa Mahalanobis-ova udaljenost (M_{ij}) od izabrane kompanije do svih jedinki iz skupa netretiranih kompanija, gdje je M_{ij} = (Z_j - Z_i)'Ω^{-1}(Z_j - Z_i). Ω je kovarijacijska matrica dobivena iz uzorka, mjereći kovarijaciju svih tretiranih jedinki sa svim netretiranima. Z_i je ili varijabla koja predstavlja vjerojatnost sklonosti ili vektor više kovarijabli (uključujući i vjerojatnost sklonosti kao jednu od njih) na temelju kojih vršimo uparivanje. Neka je ϵ unaprijed zadana maksimalna dopuštena udaljenost između Z_i i Z_j. Ako je $\|Z_j - Z_i\| > \epsilon$, taj par uklanja se iz daljnjeg razmatranja. Od parova koji su zadovoljili prethodni uvjet odabire se onaj koji ima najmanju udaljenost. Ako takvih parova nema, trenutno odabrana tretirana jedinka ne uparuje se ni sa kim, niti ona sudjeluje u procjeni. Odabrana jedinka iz kontrolne skupine ne briše se iz tog skupa kako bismo je mogli ponovno uključiti u analizu.

Gore navedeni postupak ponavlja se za sve tretirane kompanije.

Prosječan efekt tretmana tretiranih jedinki računa se kao:

$$1/n\left(\sum_{i=1}^{n} y_{1i} - \sum_{i=1}^{n} y_{0i}\right) \tag{2.4}$$

po svim parovima i.

Budući da se jedinke iz kontrolne skupine mogu pojavljivati u više različitih parova, t- statistika za testiranje jednakosti očekivanja je pristrana. Preciznije, standardne greške moraju se korigirati kako bi se mogli donijeti valjani zaključci. Lechner (vidi [6]) je 2001. godine konstruirao procjenitelja standardnih grešaka, te je taj procjenitelj primijenjen u ovom istraživanju.

2.5 Podaci

Podaci koji se koriste dolaze iz Community Innovation Survey-a (CIS) iz Flandrije, preciznije iz CIS4 za godine 2002- 2004, iz CIS5 za godine 2004- 2006 i iz CIS6 za godine 2006-2008, a dodani su i Belfirst podaci, kao i ICAROS podaci od IWT-a (Agency for Innovation by Science and Technology) koji sadrže informacije o veličini subvencija, subvencijama iz prošlih razdoblja i trajanju financiranih projekata.

Ukupan broj opservacija je 4,761, od čega su 1,948 inovativne kompanije, a 292 kompanije dobile su R&D subvenciju od flamanske vlade. Pod pojmom inovativne kompanije smatra se ona koja je u promatranom periodu implementirala novi ili značajno poboljšani proizvod, uslugu ili proces.

Tablica 2.1 (Table A1) prikazuje strukturu promatranih kompanija.

Table A.1
Industry structure.

Industry	Number of firms
Food, beverages and tobacco	364
Textiles, clothing and leather	213
Chemicals (incl. pharma), rubber and plastics	346
Metal	400
Machinery and vehicles	415
Electronics, communication and instruments	193
Other manufacturing industries	1400
Trade	738
ICT services	415
Other business services	277
Total	4761

Tablica 2.1: Table A1- izvor: D. Czarnitzki, C. Lopes-Bento, *Value for money? New microeconometric evidence on public R&D grants in Flanders*, Res. Policy (2012)

2.6 Varijable

Varijable odziva su $RDINT$ (omjer unutarnje R&D potrošnje naspram prodaje pomnožen sa 100) i $RDEMP$ (omjer R&D zaposlenosti i ukupne zaposlenosti pomnožen sa 100). Pod pojmom unutarnjeg R&D-a misli se na onaj dio koji ostaje unutar kompanije.

Kontrolnih varijabli ima nekoliko. EMP predstavlja broj zaposlenika. Promatra se $lnEMP$ i $(lnEMP)^2$. GP je dummy varijabla koja označava je li kompanija dio grupacije, a ako je, da li je njezino sjedište u inozemstvu ($FOREIGN$). Naime, kompanije koje su dio grupacije mogu imati manju motivaciju za prijavu za dobivanje subvencije budući da manje kompanije koje imaju većinsko vlasništvo nad dionicama nisu kvalificirane za SME programe s velikim iznosima subvencija. S druge strane, činjenica da su dio grupacije povećava njihovu vjerojatnost odabira jer je unutar grupacije i veća vjerojatnost efekta prelijevanja. Također, unutar grupacije je i bolja komunikacijska struktura što dovodi do bolje informiranosti o mogućnostima dobivanja subvencija. Ako je sjedište kompanije u inozemstvu, moguće je da matično poduzeće više preferira prijavu za subvenciju u svojoj zemlji, a i lokalne kompanije preferirane su pri odabiru.

Varijabla AGE predstavlja starost poduzeća, a u analizi se promatra $lnAGE$. Starije kompanije često su protiv uvođenja inovacija te se stoga ne prijavljuju za dobivanje subvencije. PS predstavlja broj patenata (*eng. patent stock*), a uključen je jer prethodni uspjesi kompanija u R&D-u povećavaju šansu za njihov odabir. U regresiju ulazi varijabla PS/EMP kako bi se uklonila potencijalna koreliranost sa veličinom kompanije.

$EXPORT$ predstavlja omjer izvoza i prodaje. Poduzeća koja puno izvoze često su i inovativnija pa time i sklonija prijavi za dobivanje subvencije. Produktivnost rada, dana varijablom LAB_PRO, uključena je jer vlada često odabire one kompanije koje imaju visoku produktivnost rada slijedeći "$picking - the - winner$" strategiju.

Nadalje, $IWT\_PAST3YRS$ označava broj dovršenih IWT projekata iz prethodne tri godine, a također su uključene i dummy varijable koje označavaju makroekonoske šokove (vremenska dummy varijabla) i dummy varijable koje označavaju kojem industrijskom sektoru kompanija pripada.

2.6.1 Vremensko usklađivanje varijabli

Svako istraživanje prati razdoblje od tri godine. Ako su zavisne varijable mjerene u vremenu t, tada su EMP, PS/EMP, LAB_PRO i $EXPORT$ mjerene na početku tog istraživanja, tj. u trenutku $t-2$. GP i $FOREIGN$ mjerene su tako da pokrivaju cijelo razdoblje od $t-2$ do t, dok je AGE egzogena varijabla i mjeri se u t.

2.7 Deskriptivna statistika

U tablici 2.2 ($Table$ 2) vidimo da su srednje vrijednosti svih varijabli značajno različite gledano između poduzeća koja su primila subvenciju od IWT-a i onih koja nisu. Iz gornjih podataka možemo zaključiti da su tretirana poduzeća u prosjeku veća, imaju veći izvoz, češće pripadaju grupaciji i imaju veću produktivnost rada. Nadalje, u posljednje tri godine imaju značajno više subvencioniranih projekata i patenata po zaposleniku. Također, $RDINT$ i $RDEMP$ su značajno veći.

2.8 Ekonometrijski rezultati

U tablici 2.3 ($Table$ 3) vidimo procjenu vjerojatnosti sklonosti, tj. procjenu vjerojatnosti dobivanja IWT subvencije dobivenu iz probit modela.

Table 2
Descriptive statistics.

Variables	Unsubsidized firms, *N* = 4469		Subsidized firms, *N* = 292		Results of *t*-tests on mean differences
	Mean	Standard deviation	Mean	Standard deviation	
Covariates					
IWT_PAST3YRS	0.023	0.181	0.736	2.436	***
*PS/EMPL*1000*	1.340	8.254	13.456	27.277	***
ln *EMP*	3.484	1.298	4.690	1.884	***
FOREIGN	0.245	0.430	0.284	0.452	
EXPORT	0.376	0.484	0.545	0.499	***
GP	0.466	0.499	0.664	0.473	***
ln *AGE*	3.113	0.777	3.153	0.889	
ln *LAB_PRO*	5.197	0.867	5.279	0.701	*
Year 2006	0.257	0.437	0.349	0.478	***
Year 2008	0.426	0.494	0.257	0.438	***
Outcome variables					
RDINT	0.894	5.322	7.579	12.694	***
RDEMP	2.646	9.960	18.287	21.980	***

Note: *** (**, *) indicate a significance level of 1% (5%, 10%)

Tablica 2.2: Table 2- izvor: D. Czarnitzki, C. Lopes-Bento, *Value for money? New microeconometric evidence on public R&D grants in Flanders*, Res. Policy (2012)

Sve varijable osim oznake grupacije (GP), starosti ($lnAGE$) i produktivnosti rada ($lnLAB\_PRO$) su značajne za procjenu vjerojatnosti, tj. koeficijenti uz njih su značajno različiti od nule na razini značajnosti od 1%. $lnEMP$ nije značajna, no njezin kvadratni oblik jest. McFadden R^2 je 0.308 što nam govori da je fit modela dobar.

14 opservacija iz skupa tretiranih kompanija uklanja se iz analize jer su njihove vjerojatnosti sklonosti veće od maksimalne ili manje od minimalne vrijednosti vjerojatnosti sklonosti kontrolne skupine. Dodatnih 26 tretiranih kompanija uklanja se zbog uvjeta maksimalne udaljenosti između jedinki koje se uparuju. Ukupno se dakle briše 40 kompanija iz skupa tretiranih te u tome skupu ostaju 252 kompanije. Rezultati metode uparivanja za reducirani skup dani su u tablici 2.4 ($Table$ 4):

p-vrijednosti t-testa o jednakosti očekivanja za sve varijable veće su od 0.1, osim za $RDINT$ i $RDEMP$. To nam govori da je uparivanje bilo uspješno. Kod varijabli $RDINT$ i $RDEMP$ postoji značajna razlika između kontrolne i tretirane skupine i tu razliku možemo

Table 3
Probit estimation.

	Coefficient	Standard errors
IWT_PAST3YRS	0.755***	0.084
*PS/EMPL*1000*	0.016***	0.002
ln *EMP*	-0.140	0.105
(ln *EMP*)^2	0.042***	0.011
FOREIGN	-0.429***	0.099
EXPORT	0.655***	0.103
GP	0.098	0.091
ln *AGE*	-0.073	0.047
ln *LAB_PRO*	0.008	0.052
Intercept	-1.750***	0.407
Test on joint significance of industry dummies		$\chi^2(9) = 94.29$***
Test on joint significance of time dummies		$\chi^2(2) = 24.61$***
McFadden R^2		0.308
Number of observations		4761

Note : *** (**, *) indicate a significance level of 1% (5%, 10%)

Tablica 2.3: Table 3- izvor: D. Czarnitzki, C. Lopes-Bento, *Value for money? New micro-econometric evidence on public R&D grants in Flanders*, Res. Policy (2012)

pripisati djelovanju subvencije. Prosječan efekt tretmana gledano za RDINT je i do 3.73%, a za RDEMP i do 9.57%. Budući da primanje subvencije pozitivno utječe na te dvije varijable, možemo odbaciti pretpostavku o potpunom efektu istiskivanja.

U skladu s prijašnjim saznanjima, možemo zaključiti da je utjecaj IWT subvencija okidač za investiranje u *R&D* u kompaniji koja prima tu subvenciju.

2.8.1 Stabilnost efekta tretmana tokom vremena

Sada ćemo promatrati mijenja li se efekt tretmana tokom vremena. Gledat ćemo podatke za godine 2004., 2006. i 2008. kako bismo vidjeli mijenja li se utjecaj primanja subvencije u periodima od t, t+1, do t+2. Rađena je regresija efekta tretmana na dummy varijable Y2006 i Y2008 (2004 je bazna godina). Y2006 i Y2008 označavaju je li kompanija primila subvenciju u godinama 2006., odnosno 2008. Kako bi efekt tretmana bio stabilan kroz

Table 4 Matching results on the full sample.					
Variables	Selected control group, *N* =252		Subsidized firms, *N* =252		*p* -Value of *t* -tests on mean difference
	Mean	Standard deviation	Mean	Standard deviation	
Covariates					
IWT_PAST3YRS	0.258	0.709	0.349	0.821	$p = 0.222$
*PS/EMPL*1000*	7.932	20.092	8.354	20.032	$p = 0.831$
ln *EMP*	4.302	1.876	4.359	1.678	$p = 0.745$
$(\ln EMP)^2$	22.011	17.906	21.808	15.586	$p = 0.904$
FOREIGN	0.262	0.441	0.262	0.441	$p = 1.000$
EXPO	0.599	0.491	0.560	0.497	$p = 0.415$
GP	0.544	0.499	0.623	0.486	$p = 0.103$
ln *AGE*	3.099	0.873	3.085	0.860	$p = 0.866$
ln *LAB_PRO*	5.291	0.792	5.252	0.701	$p = 0.606$
Outcome variables					
RDINT	3.151	10.373	6.883	12.140	$p = 0.001$
RDEMP	8.061	18.254	17.627	21.603	$p < 0.001$

Tablica 2.4: Table 4- izvor: D. Czarnitzki, C. Lopes-Bento, *Value for money? New micro-econometric evidence on public R&D grants in Flanders*, Res. Policy (2012)

vrijeme, niti jedna od dummy varijabli ne bi trebala biti značajno različita od nule niti bi cijela regresija trebala biti značajna. Iz tablice 2.5 (*Table* 5) vidimo da su i za razinu značajnosti $\alpha = 10\%$ obje p-vrijednosti veće od α pa ne možemo odbaciti hipotezu da su koeficijenti uz Y2006 i Y2008 jednaki nuli. Stoga zaključujemo da je efekt IWT subvencije na *R&D* zaposlenost i investicije stabilan, tj. ne opada tijekom vremena.

2.8.2 Učinak više subvencioniranih projekata istovremeno

Zanima nas je li efekt tretmana veći ukoliko kompanija ima više subvencioniranih projekata istovremeno. Naime, u promatranom uzorku prosječan broj projekata po kompaniji je 1.5. Od 292 tretirane kompanije, 66% kompanija ima samo jedan projekt, 19% njih ima dva projekta, a jedna kompanija ima čak 26 projekata.

U tablici 2.6 (*Table* 6) vidimo rezultate regresije efekta tretmana na varijablu koja predstavlja broj IWT projekata unutar kompanije. Za svaku razinu značajnosti α je p-vrijednost

Table 5
OLS regression testing for stability of treatment effect over time.

	Dependent variable: α_i					
	R&D employment			R&D intensity		
	Coefficient	Standard errors	$p > \|t\|$	Coefficient	Standard errors	$p > \|t\|$
Y2006	5.299	3.985	0.185	0.905	1.824	0.620
Y2008	1.513	4.578	0.741	-0.017	2.988	0.995
Intercept	7.317	3.285	0.027	3.424	1.619	0.036
Number of observations	252			252		
Overall significance	$F(2, 208) = 0.16$			$F(2, 208) = 1.02$		

Note : Standard errors are clustered at the firm level, as a few firms appear more than once in the database. *** (**, *) indicate a significance level of 1% (5%, 10%).

Tablica 2.5: Table 5- izvor: D. Czarnitzki, C. Lopes-Bento, *Value for money? New micro-econometric evidence on public R&D grants in Flanders*, Res. Policy (2012)

Table 6
Regression on the treatment effect on the number of supported procjects.

	Dependent variable: α_i					
	R&D employment			R&D intensity		
	Coefficient	Standard errors	$p > \|t\|$	Coefficient	Standard errors	$p > \|t\|$
Number of IWT projects	2.321	0.521	0.000	0.916	0.306	0.003
Intercept	5.789	2.122	0.007	2.240	1.141	0.051
Number of observations	252			252		
Overall significance	$F(1, 208) = 19.83$***			$F(1, 208) = 8.96$***		

Note : Standard errors are clustered at the firm level, as a few firms appear more than once in the database. *** (**, *) indicate a significance level of 1% (5%, 10%).

Tablica 2.6: Table 6- izvor: D. Czarnitzki, C. Lopes-Bento, *Value for money? New micro-econometric evidence on public R&D grants in Flanders*, Res. Policy (2012)

$p \leq \alpha$ pa odbacujemo hipotezu da je koeficijent uz broj IWT projekata jednak nuli. Također, cijela regresija je značajna, tj. zaključujemo da reducirani model nije dovoljan. Koeficijent uz broj IWT projekata je jednak 2.321, tj. pozitivan je pa možemo zaključiti da veći broj projekata pojačava efekt tretmana.

2.8.3 Učinak uzastopnog ponavljanja subvencija na efekt tretmana

Provođenjem regresije efekta tretmana na dummy varijablu koja označava je li u prethodne 3 godine kompanija primila subvenciju želi se procijeniti je li poželjno ponavljati davanje subvencije istoj kompaniji. Iz tablice 2.7 (*Table* 7) vidimo da efekt tretmana nije manji ako se uzastopno daje subvencija. Naime, zbog velike p-vrijednosti ne možemo odbaciti hipotezu da je koeficijent uz dummy varijablu jednak nuli, tj. dummy varijabla nema utjecaja na efekt tretmana.

Table 7
Regression testing the treatment effect of "consecutive clients".

	Dependent variable: α_i					
	R&D employment			R&D intensity		
	Coefficient	Standard errors	$p > \|t\|$	Coefficient	Standard errors	$p > \|t\|$
Dummy (IWTsubsidy in the last 3 years)	-5.043	7.768	0.517	4.581	3.503	0.192
Intercept	9.927	1.814	0.000	3.151	1.054	0.001
Number of observations	252			252		
Overall significance	$F(1, 208) = 0.42$			$F(1, 208) = 1.71$		

Note : Standard errors are clustered at the firm level, as a few firms appear more than once in the database. *** (**, *) indicate a significance level of 1% (5%, 10%).

Tablica 2.7: Table 7- izvor: D. Czarnitzki, C. Lopes-Bento, *Value for money? New microeconometric evidence on public R&D grants in Flanders*, Res. Policy (2012)

2.8.4 Učinak primanja subvencija iz drugih izvora na efekt tretmana

Kako bi se testirao utjecaj primanja subvencija iz drugih izvora, potrebno je promijenit način uparivanja jedinki. Umjesto uparivanja samo na temelju vjerojatnosti sklonosti, za parove će nužno vrijediti sljedeća dva uvjeta:

1. jedan član para je primio IWT subvenciju, a drugi nije,

2. ili će za obje jedinke iz para vrijediti da su primile dodatnu subvenciju (iz drugog izvora) ili će vrijediti da je niti jedna od njih nije primila.

Iz tablice 2.8 ($Table$ 8) vidimo da su jedinke dobro uparene jer nema značajne razlike između tretirane i kontrolne skupine. Razliku za $RDINT$ i $RDEMP$ opet možemo pripisati djelovanju subvencije. Rezultate regresije efekta tretmana na dummy varijablu koja označava primanje subvencije iz drugih izvora možemo vidjeti u tablici 2.9 ($Table$ 9). Procijenjeni koeficijent uz dummy varijablu je negativan, no zbog velike p-vrijednosti ne možemo odbaciti hipotezu da je taj koeficijent jednak nuli. Zaključujemo da dodatne subvencije ne utječu na efekt tretmana pa time ni na efekt istiskivanja.

Table 8
Matching results, full sample: controlling for other subsidies.

Variables	Selected control group, N = 215		Subsidized firms, N = 215		p-value of t-tests on mean differences
	Mean	Standard deviation	Mean	Standard deviation	
Covariates					
IWT_PAST3YRS	0.233	0.613	0.298	0.680	$p = 0.350$
*PS/EMPL*1000*	5.628	17.357	6.191	17.597	$p = 0.767$
ln *EMP*	4.482	1.927	4.318	1.656	$p = 0.337$
$(\ln EMP)^2$	23.794	18.958	21.376	15.194	$p = 0.174$
FOREIGN	0.293	0.456	0.247	0.432	$p = 0.339$
EXPO	0.553	0.498	0.540	0.500	$p = 0.932$
GP	0.619	0.487	0.623	0.486	$p = 0.930$
ln *AGE*	3.148	0.851	3.075	0.852	$p = 0.465$
ln *LAB_PRO*	5.331	0.739	5.230	0.701	$p = 0.387$
Outcome variables					
RDINT	3.864	10.655	6.780	12.171	$p = 0.015$
RDEMP	9.633	18.772	17.635	21.746	$p < 0.001$

Tablica 2.8: Table 8- izvor: D. Czarnitzki, C. Lopes-Bento, *Value for money? New microeconometric evidence on public R&D grants in Flanders*, Res. Policy (2012)

2.8.5 Promatranje inovativnih kompanija zasebno

Promatraju se samo inovativne kompanije kako bi se vidjelo ovisi li efekt tretmana u velikoj mjeri o neinovatorima iz kontrolne skupine. Prvi korak je probit procjena za novi skup - skup u kojem su samo inovativne kompanije. Rezultate te procjene vidimo u tablici 2.10 ($Table$ 10) te zaključujemo da su značajne iste varijable kao i kod probit procjene na cijelom skupu.

Table 9
Regression of treatment effect on the receipt of other subsidies (215 observations).

	R&D employment			R&D intensity		
	Coefficient	Standard errors	$p > \|t\|$	Coefficient	Standard errors	$p > \|t\|$
Dummy (subsidies received from other entities than the IWT)	-3.564	4.101	0.386	-0.034	2.161	0.987
Intercept	9.460	2.207	0.000	2.929	1.331	0.029
Number of observations	215			215		
Overall significance	$F(1, 183) = 0.76$			$F(1, 183) = 0.01$		

Note : Standard errors are clustered at the firm level, as a few firms appear more than once in the database.

Tablica 2.9: Table 9- izvor: D. Czarnitzki, C. Lopes-Bento, *Value for money? New microeconometric evidence on public R&D grants in Flanders*, Res. Policy (2012)

Drugi korak je uparivanje jedinki. Iz tablice 2.11 ($Table$ 11) vidljivi su rezultati tog uparivanja. p- vrijednosti su dosta velike pa nema značajne razlike u prosječnim vrijednostima varijabli između kontrolne i tretirane skupine, za sve varijable osim za $RDINT$ i $RDEMP$. Rezultati za $RDINT$ i $RDEMP$ slični su onima iz tablice 2.4 ($Table$ 4). To nam govori da pri uparivanju tretiranih kompanija sa čitavom kontrolnom skupinom (onom u kojoj su i inovativne i neinovativne kompanije), većinu parova ionako čine inovatori. Zaključujemo da efekt tretmana ne ovisi u velikoj mjeri o neinovatorima iz kontrolne skupine.

2.8.6 Promatranje KMO programa zasebno

Oko dvije trećine IWT subvencija u razdoblju od 2004. do 2010. godine odnosilo se na poticanje R&D-a i i inovacija u malim i srednje velikim poduzećima. Takav tip subvencija naziva se "KMO program". Stoga se sada promatra skup u kojem se nalaze samo kompanije koje su podobne za primanje KMO subvencije. Probit procjena na tom skupu daje slične rezultate kao i procjena na čitavom skupu pa ovdje ne navodimo tablicu s rezultatima.

U tablici 2.12 ($Table$ 12) vidimo rezultate uparivanja. Kao i prije, p-vrijednosti su dosta velike pa zaključujemo da je uparivanje dobro provedeno. Rezultati su slični onima iz tablice 2.4 ($Table$ 4) kada smo promatrali sve IWT subvencije. Ipak, vidimo da je efekt tretmana nešto veći kod KMO subvencija, što je i očekivano budući da davanje subvencije manjoj kompaniji ima veći učinak na R&D nego što je to slučaj kod većih kompanija.

Table 10
Probit estimation, only for innovative firms.

	Coefficient	Standard errors
IWT_PAST3YRS	0.648***	0.084
*PS/EMPL*1000*	0.014***	0.002
ln *EMP*	-0.158	0.118
(ln *EMP*)^2	0.038***	0.012
FOREIGN	-0.423***	0.109
EXPO	0.442***	0.128
GP	0.072	0.103
ln *AGE*	-0.070	0.052
ln *LAB_PRO*	0.006	0.063
Intercept	-1.339***	0.488
Test on joint significance of industry dummies	χ^2 (9) = 64.04***	
Test on joint significance of time dummies	χ^2 (2) = 15.16***	
Number of observations	1.948	

Note : Standard errors are clustered at the firm level, as a few firms appear more than once in the database. *** (**, *) indicate a significance level of 1% (5%, 10%)

Tablica 2.10: Table10- izvor: D. Czarnitzki, C. Lopes-Bento, *Value for money? New microeconometric evidence on public R&D grants in Flanders*, Res. Policy (2012)

Također, KMO subvencije definirane su tako da pokrivaju 10% više troškova R&D projekata od drugih tipova subvencija.

2.8.7 Test robusnosti pomoću instrumentalnih varijabli

Korištenje metode uparivanja pretpostavljalo je da su u model uključene sve varijable koje imaju utjecaj na primanje IWT subvencije. Za procjenu efekta tretmana mogu se pokušati pronaći i instrumentalne varijable za tretman te pomoću njih napraviti procjena.

Kao što smo već prije spomenuli, za SME kompanije vrijede drukčija pravila subvencioniranja. Te kompanije podobne su za prijavu za KMO programe i općenito primaju veće iznose za pokrivanje troškova projekta. Pod uvjetom da nisu članovi grupacije, srednje velikim kompanijama pokriva se i do 35% troška, dok se manjima pokriva i do 45%. Stoga bi potencijalan instrument mogla biti dummy varijabla koja je jednaka 1 ako je kompanija

Table 11
Matching results, of innovative firms only.

Variables	Selected control group, N = 262		Subsidized firms, N = 262		p-Value on the t-test on mean difference
	Mean	Standard deviations	Mean	Standard deviations	
Covariates					
IWT_PAST3YRS	0.305	0.726	0.321	0.766	$p = 0.839$
*PS/EMPL*1000*	10.053	23.261	10.500	23.534	$p = 0.851$
ln *EMP*	4.387	1.648	4.426	1.707	$p = 0.818$
$(\ln EMP)^2$	21.949	15.291	22.490	16.121	$p = 0.733$
FOREIGN	0.271	0.445	0.263	0.441	$p = 0.866$
EXPORT	0.595	0.492	0.573	0.496	$p = 0.648$
GP	0.595	0.492	0.634	0.483	$p = 0.442$
ln *AGE*	3.149	0.855	3.091	0.860	$p = 0.507$
ln *LAB_PRO*	5.164	0.714	5.260	0.708	$p = 0.188$
Outcome variables					
RDINT	3.235	8.328	6.745	11.622	$p < 0.001$
RDEMP	8.145	13.390	17.220	20.670	$p < 0.001$

Tablica 2.11: Table11- izvor: D. Czarnitzki, C. Lopes-Bento, *Value for money? New microeconometric evidence on public R&D grants in Flanders*, Res. Policy (2012)

mala i ne pripada grupaciji, a analogno bismo mogli konstruirati dummy varijablu i za srednje velike kompanije. Naime, kompanije podobne za KMO programe imaju veći poticaj za prijavu budući da je za isti trošak prijave, njihov potencijalni dobitak veći. Također, tako konstruirana varijabla je korelirana sa endogenom dummy varijablom koja označava tretman i nekorelirana sa ostalim varijablama- uzimajući u obzir da definiciju male, tj. srednje velike kompanije određuje Europska komisija i da sama kompanija ne određuje broj zaposlenih na način da bolje pogoduje nekom kriteriju programa subvencioniranja.

Nažalost, u prvom koraku regresije na tretman (tj. na primanje subvencije) niti jedna od tih dviju dummy varijabli nije značajna. Stoga je potrebno pronaći drugi potencijalni instrument.

Za instrumente se može uzeti broj subvencioniranih projekata koji su završili u trenutku t-2, skupa sa njihovom prosječnom veličinom (ukupan iznos subvencije u tisućama EUR-a podijeljen sa brojem subvencioniranih projekata). U prvom koraku regresije na trenutan broj subvencioniranih projekata, oba instrumenta su značajna. Vrijednosti Hansenovog J testa su male što govori da je koreliranost instrumenata sa greškom modela mala. U tablici 2.13 ($Table$ 13) vidimo rezultate regresije koja pokazuje odnos trenutnog broja subvenci-

Table 12
Matching results, KMO recipients only

Variables	Selected control group, N = 112		Subsidized firms, N = 112		p-Value on the t-test on mean difference
	Mean	Standard deviations	Mean	Standard deviations	
Covariates					
KMO_PAST3YRS	0.098	0.354	0.152	0.429	p = 0.327
*PS/EMPL*1000*	5.672	16.376	5.873	16.980	p = 0.536
ln *EMP*	3.355	1.093	3.282	1.027	p = 0.626
$(\ln EMP)^2$	12.439	6.815	11.816	6.379	p = 0.502
FOREIGN	0.045	0.207	0.036	0.186	p = 0.748
EXPORT	0.616	0.488	0.553	0.499	p = 0.366
GP	0.429	0.497	0.330	0.472	p = 0.149
ln *AGE*	2.884	0.771	2.995	0.797	p = 0.311
ln *LAB_PRO*	5.123	0.683	4.999	0.617	p = 0.178
Outcome variables					
RDINT	2.583	7.323	7.088	12.780	p = 0.002
RDEMP	8.533	20.722	18.726	22.678	p = 0.001

Tablica 2.12: Table12- izvor: D. Czarnitzki, C. Lopes-Bento, *Value for money? New microeconometric evidence on public R&D grants in Flanders*, Res. Policy (2012)

oniranih projekata sa R&D intenzitetom i zaposlenosti. Promatra se cijeli uzorak te uzorak inovativnih kompanija zasebno.

Vidimo da je broj subvencioniranih projekata značajan i ima pozitivan efekt. Veličina efekta je slična rezultatima dobivenima pomoću metode uparivanja. Time su potvrđeni rezultati dobiveni metodom uparivanja, tj. može se zaključiti da je model dobar.

Table 13
Instrumental variable regressions using full sample and subsample of innovators.

Variables	IV regression, full sample (4,761 observations)				IV regression, innovator sample (1,948 observations)			
	R&D INTENSITY		*R&D EMPLOYMENT*		*R&D INTENSITY*		*R&D EMPLOYMENT*	
	Model 1		Model 2		Model 3		Model 4	
# of current IWT projects	1.242	***	2.701	***	1.074	***	2.342	***
	(0.369)		(0.629)		(0.314)		(0.508)	
*PS/(EMPL*1000)*	107.434	***	203.829	***	113.718	***	202.138	***
	(25.848)		(41.061)		(27.862)		(41.875)	
ln *EMP*	0.407		-1.104		0.162		-4.152	**
	(0.381)		(1.058)		(0.746)		(1.657)	
(ln *EMP*)^2	-0.064		0.044		-0.063		0.24	
	(0.049)		(0.120)		(0.082)		(0.168)	
FOREIGN	0.469		0.103		1.298	*	1.152	
	(0.374)		(0.557)		(0.720)		(1.018)	
EXPORT	0.827	***	2.71	***	1.218	**	4.073	***
	(0.197)		(0.378)		(0.587)		(0.966)	
GP	0.516	**	0.784	*	1.211	**	1.091	
	(0.250)		(0.414)		(0.538)		(0.856)	
ln *AGE*	-0.2	**	-0.533	**	-0.22		-0.636	
	(0.095)		(0.209)		(0.193)		(0.417)	
ln *LAB_PRO*	-0.446	***	0.109		-1.155	***	0.499	
	(0.133)		(0.185)		(0.382)		(0.501)	
Intercept	2.834	***	7.419	***	7.447	***	16.171	***
	(0.971)		(2.490)		(2.653)		(5.130)	
F test of excluded instruments	32.92	***	32.92	***	33.17	***	33.17	***
Hansen *J* over identification test	2.14		1.486		2.424		1.921	

Notes : *** (**, *) indicate a significance level of 1% (5%, 10%). Standard errors in parentheses are clustered. All models contain full sets od industry and time dummies (not reported). The number of current IWT projects has been instrumented with the number of IWT projects that ended in period *t* -2 as well as the average size (in terms of thsd. EUR) per project.

Tablica 2.13: Table13- izvor: D. Czarnitzki, C. Lopes-Bento, *Value for money? New microeconometric evidence on public R&D grants in Flanders*, Res. Policy (2012)

Bibliografija

[1] J. D. Angrist, *Instrumental Variables Estimation of Average Treatment Effects in Econometriccs and Epidemiology*, National Bureau of Economic Research Technical Working Paper Number 115 (1991).

[2] J. D. Angrist, G. W. Imbens, *Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity*, Journal of the American Statistical Association 90 (1995), 431- 442.

[3] K.J. Arrow, *Economic welfare and the allocation of resources for invention. In: Nelson, R.R. (Ed.), The Rate and Direction of Inventive Activity: Economic and Social Factors*, National Bureau of Economic Research, Conference Series. Princeton University Press, Princeton (1962), pp. 609–625.

[4] W. Cochran, D. B. Rubin, *Controlling bias in observational studies*, Sankhya 35 (1973), 417–446.

[5] D. Czarnitzki, C. Lopes-Bento, *Value for money? New microeconometric evidence on public R&D grants in Flanders*, Res. Policy (2012), http://dx.doi.org/10.1016/j.respol.2012.04.008.

[6] M. Lechner, *Identification and estimation of causal effects of multiple treatments under the conditional independence assumption*, In: Lechner, M., Pfeiffer, F. (Eds.), Econometric Evaluation of Labour Market Policies. Physica, Heidelberg (2001), pp. 43–58.

[7] P. R. Rosenbaum, D. B. Rubin, *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, Biometrika 70 (1983), 41- 55.

[8] J. M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, 2002.

Sažetak

U ovome radu koncentrirali smo se na procjenu prosječnog efekta tretmana koristeći različite pretpostavke i ovisno o njima regresijske metode, metode bazirane na vjerojatnosti sklonosti ili metode instrumentalnih varijabli. Varijable y_1 i y_0 označavale su ishod sa, odnosno bez tretmana, dok je binarna varijabla w označavala tretman, a $\mathbf{x}$ vektor opaženih kovarijabli.

Ono što smo htjeli procijeniti bila je razlika između ishoda sa i bez tretmana, y_1- y_0. Budući da se pojedinac ne može istovremeno nalaziti u oba stanja, tu razliku nije moguće direktno izračunati. Stoga smo promatrali sljedeća dva očekivanja: **prosječan efekt tretmana (average treatment effect - ATE)** koji je definiran kao: $ATE \equiv E(\ y_1 - y_0\ )$ i **prosječan efekt tretmana tretiranih jedinki (average treatment effect on the treated -** ATE_1**)** definiran sa: $ATE_1 \equiv E(\ y_1 - y_0 \mid w = 1)$. Također, promatrali smo i **prosječan efekt tretmana uvjetno na x** i **prosječan efekt tretmana tretiranih uvjetno na x**: $ATE_1(\mathbf{x}) = E(\ y_1 - y_0 \mid \mathbf{x}, w = 1)$ i $E(\ y_1 - y_0 \mid \mathbf{x}) = ATE(\mathbf{x})$.

Ključna pretpostavka bila je **nezavisnost o tretmanu** (*eng. ignorability of a treatment*) dana uvjetima:

ATE.1: Uvjetno na $\mathbf{x}$, w i $(\ y_0, y_1)$ su nezavisni.

ATE.1': (a) $\mathrm{E}(y_0 \mid \mathbf{x}, w) = \mathrm{E}(y_0 \mid \mathbf{x})$; (b) $\mathrm{E}(y_1 \mid \mathbf{x}, w) = \mathrm{E}(y_1 \mid \mathbf{x})$.

Propensity score ili **vjerojatnost sklonosti** $p(\mathbf{x})$ označava vjerojatnost da jedinka primi tretman ako nam je poznat vektor kovarijabli $\mathbf{x}$. Važna pretpostavka pri korištenju metoda baziranih na vjerojatnosti sklonosti bila je **jaka nezavisnost o tretmanu** (*eng. strong ignorability of a treatment*) (uz danu kovarijablu $\mathbf{x}$) koja se sastoji od pretpostavke ATE.1 i uvjeta $0 < p(\mathbf{x}) < 1,\ \forall\ \mathbf{x}$. Koristeći te metode objasnili smo ideju **uparivanja na temelju vjerojatnosti sklonosti**, koju su prvi put predložili Rosenbaum i Rubin 1983. godine.

Metoda instrumentalnih varijabli (IV) koristi se za procjenu ATE i ATE_1 kada nismo sigurni vrijedi li nezavisnost o tretmanu, ali i za procjenu **lokalnog prosječnog efekta tretmana (LATE)**. Promatrali smo najjednostavniji slučaj - kada je instrumentalna varija-

bla z binarna - te smo LATE definirali na subpopulaciji sastavljenoj od jedinki kod kojih varijabla z određuje hoće li te jedinke sudjelovati u tretmanu.

U drugom dijelu rada pokazali smo korištenje uparivanja na temelju vjerojatnosti sklonosti u praksi. Procjenjivali smo utjecaj IWT subvencija na razvoj R&D-a u kompanijama u Flandriji i kako se taj utjecaj mijenja ovisno o određenim parametrima.

Summary

In this paper focus was on estimating average treatment effects. We used different assumptions and different methods- regression methods, methods based on the propensity score and instrumental variables methods. Variable y_1 denoted the outcome with treatment and y_0 without the treatment, while $\mathbf{x}$ denoted a vector of observed covariates and w denoted receiving the treatment.

To estimate the difference y_1 - y_0, we observed **average treatment effect (ATE)** defined as $ATE \equiv E(\, y_1 - y_0 \,)$ and **average treatment effect on the treated** **(**ATE_1**)** defined as $ATE_1 \equiv E(\, y_1 - y_0 \mid w = 1)$.

Key assumption was assumption called **ignorability of treatment** (given observed covariates x):

ATE.1: Conditional on $\mathbf{x}$, w and $(\, y_0, y_1)$ are independent.

ATE.1': (a) $\mathrm{E}(y_0 \mid \mathbf{x}, w) = \mathrm{E}(y_0 \mid \mathbf{x})$; (b) $\mathrm{E}(y_1 \mid \mathbf{x}, w) = \mathrm{E}(y_1 \mid \mathbf{x})$.

Propensity score $p(\mathbf{x})$ is the probability of receiving the treatment given the covariates. For using methods based on the propensity score we needed **strong ignorability of a treatment** assumption. We described matching algorithm based on the propensity score.

For estimating the **local average treatment effect (LATE)**, we used the **instrumental variables method** in the simplest scenario- when an instrumental variable z is binary.

Finally, we showed how to use the propensity score matching in practice. We estimated the impact of IWT subsidies on firms' R&D intensity and employment in the Flemish region.